Universitat
Pompeu Fabra
Barcelona

THE bHLH PROJECT
OLIG2

Denis Expósito
Gary J. Espitia S
Marko Ludaic
Nikola Panajotovikj

CONTENT

1. INTRODUCTION

Oligodendrocyte transcription factor (OLIG2) is a basic helix-loop-helix (bHLH) transcription factor. It is encoded by the *OLIG2 gene*. The protein is of 329 amino acids in length and 32 kDa in size. OLIG2 is mostly expressed in central nervous system (CNS). It acts as anti-neurigenic and a neurigenic factor at different stages of development and it has an important role for determining motor neuron and oligodendrocyte differentitation.

By increasing the conduction velocity of axons

BIOLOGICAL FUNCTIONS

By increasing the conduction velocity of axons, myelin allows for increased body size, rapid movement and a large and complex brain. In the CNS, oligodendrocytes are the myelin-forming cells. The transcription factors OLIG1 and OLIG2, master regulators of oligodendrocyte development, presumably also played a seminal role during evolution of the genetic programme leading to myelination in the CNS.

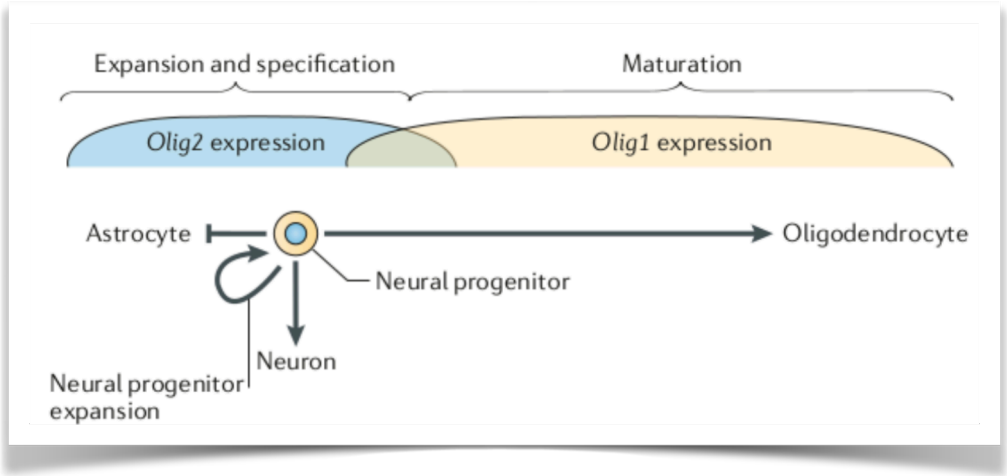


Figure 1. *Representation of OLIG2 role in astrocytes and maturation of oligodendrocytes*

OLIG2 is expressed in the ventral spinal cord as early as 9.5 dpc (days post coitum). Expression becomes progressively restricted to a narrow zone within the ventral neuroepithelium of the spinal cord. In the 14.5 dpc spinal cord, expressed in the oligodendrocyte progenitors of the ventral ventricular zone, but not dorsal root ganglia Schwann cells. Also expressed scattered in the mantle zone, likely corresponding to oligodendrocyte progenitors migrating out from their site of origin. In the brain, from 10.5 through 14.5 dpc, expressed in numerous cells in the ventricular and subventricular zones of the lateral and medial ganglionic eminences, suggesting that expression might not be limited to the oligodendrocytic lineage. By 15.5 dpc, dispersed throughout the gray matter, with little or no residual expression in the ventricular zone. In the postnatal brain, present preferentially in the white matter, such as corpus callosum and cerebellar medulla. Expressed in the 13.5 and 14.5 dpc retina and in the olfactory epithelium from 11.5 dpc onward.

Regardless of where the primary selection came from, once the myelinating programme started to evolve in one cell type, all or part of the programme could have been activated in other cells, given appropriate cues. Therefore, evolution of CNS and PNS myelin would have gone largely hand in hand. In the CNS, oligodendrocyte precursors (OLPs) acquired the ability to migrate and myelinate widely and thereby myelin became ubiquitous throughout the CNS, with major advantages for all kinds of neural processes and tasks.

2. THE USE OF dN/dS FOR THE CASE OF HUMAN OLIG2 PARALOGS

INTRODUCTION

Evolutionary pressures on proteins are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites. The dN/dS ratio was originally developed for application to distantly diverged sequences, the differences among which represent substitutions that have fixed along independent lineages (1). Positive selection can be inferred whenever the estimated ratio (v) of non-synonymous (b) to synonymous (a) substitution rates significantly exceeds one. Combining a set of results from individual sites to draw conclusions about a whole gene while controlling the false discovery rate leads to an unavoidable drop in power to detect gene-wide selection, especially when the number of taxa (which drive signal at individual sites) is limited. A branch-site unrestricted statistical test for episodic diversification (BUSTED) is capable of detecting positive selection that has acted on a subset of branches in a phylogeny at a subset of sites within the gene (2). Most current computational methods designed to detect the imprint of natural selection at a site in a protein coding gene assume the strength and direction of natural selection is constant across all lineages. The imprint of natural selection on protein coding genes is often difficult to identify because selection is frequently transient or episodic, i.e. it affects only a subset of lineages and may fail to recognize such sites, therefore a large proportion of positively selected sites. The mixed effects model of evolution (MEME) can detect adaptive evolution, even when the selective forces are not constant across taxa. MEME is capable of identifying instances of both episodic and pervasive positive selection at the level of an individual site. MEME allows the distribution of v to vary from site to site (the fixed effect) and also from branch to branch at a site (3). MEME can reliably capture the molecular footprints of both episodic and pervasive positive selection, a task for which current models are not well suited. In this study we want to evaluate positive diversifying selection for OLIG2 human paralog genes through the use of BUSTED and MEME.

METHODOLOGY

In our study of dN/dS we used the Ensembl database to extract the paralog .fas CDS files of the human gene OLIG2. The information was aligned and cleared from stop codons using the command line tool Hyphy, and described in the discussion of HyPhy github (2,4). In our case we used 16 paralogs with great diversity of amino acids (aa) range, going from 152aa up to 383aa. Then the processed data was evaluated for dN/dS using BUSTED firstly to determine the presence of positive diversifying selection in the whole gene. Then we used MEME to determine which specific regions were related. This information then was extracted and using the p-value of the authors of the MEME test (p-value=0.05), we determined a cut-off point for statistically significant likelihood ratio (LR). For LRT in MEME to work properly the alternative hypothesis and the null hypothesis have to be simple type hypotheses. The LRT is the likelihood of samples given that the alternative hypothesis is true, over the likelihood of samples given that the null hypothesis is true. In this case LRT is used to establish the likelihood of a site having a diversifying selection.

RESULTS

The results from the analysis for dN/dS are hosted and available in the Datamonkey database, links available in the appendix. Since BUSTED is a gene level method and not a site level method, its usage served as an initial evaluation for the need of other tools, such as MEME, and to determine the positive selection at site level. In our case, there was evidence of episodic diversifying selection

in this dataset (p=1.665e-16), therefore MEME was also used. The BUSTED tool also provided information of the estimated ratio (ER) (site level LR) of specific sites.

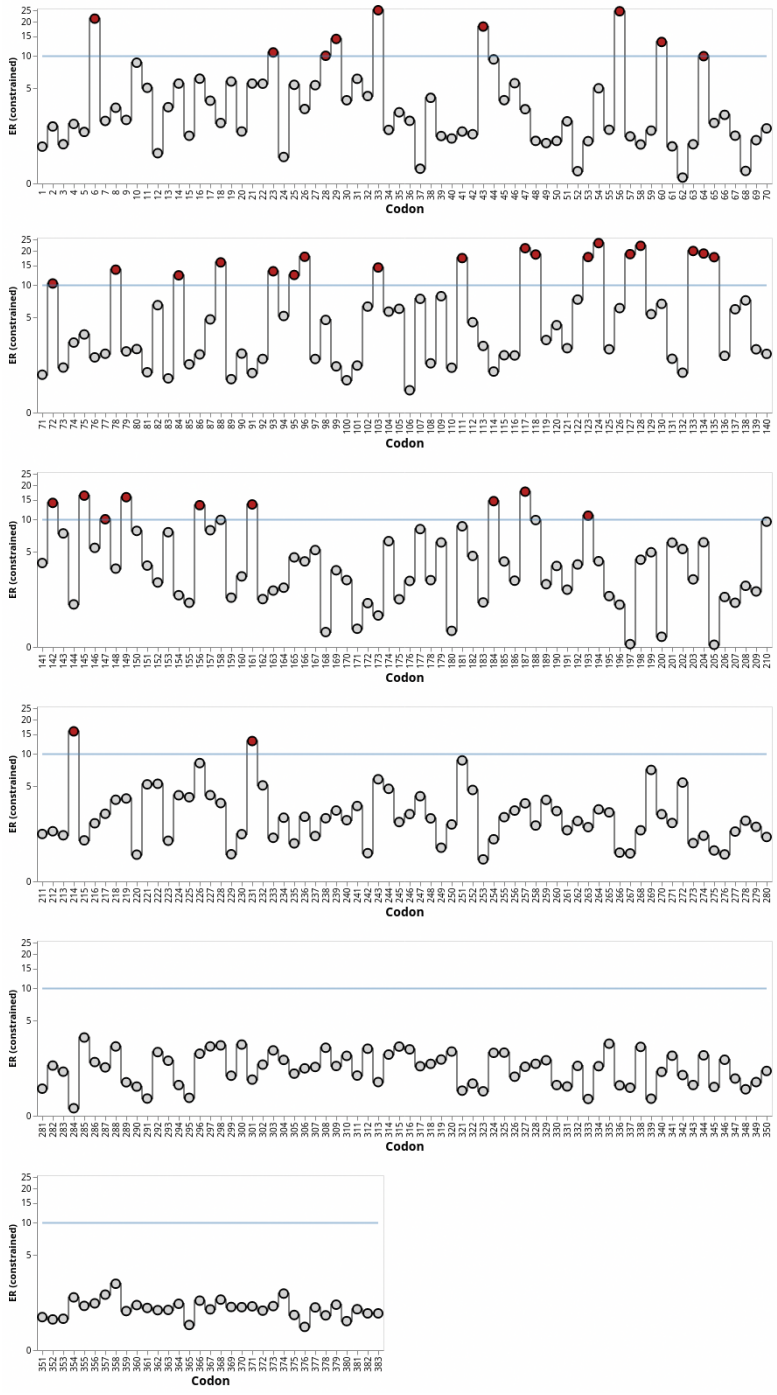


Figure 2. Evidence ratios (site level likelihood ratios) for ω>1, comparing the unrestricted model with the model where max(ω) := 1, and all other parameters are kept at their maximum likelihood values. Solid line = user selected significance threshold.

We also determined that the average cumulative LRT of 71 gives a whole gene average dN/dS 0.46 for the paralog of 152aa (ENSP00000362777) and 0.18 for the paralog of 383aa (ENSP00000318799).

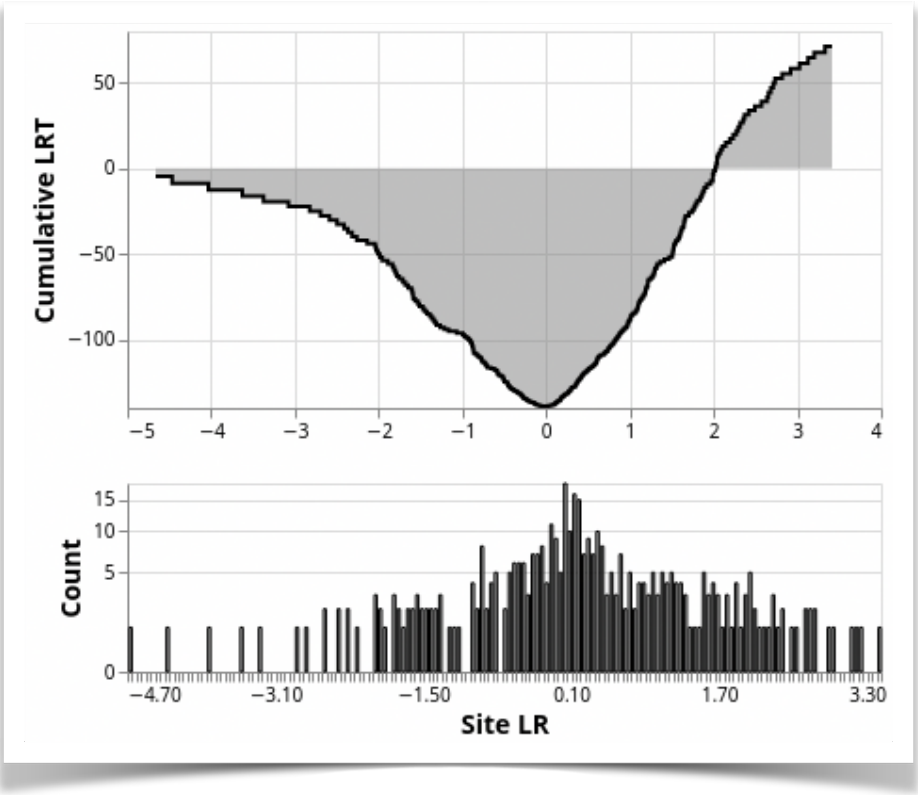


Figure 3. Cumulative distribution of the likelihood ratio test for the BUSTED test broken down by the contributions of individual sites.

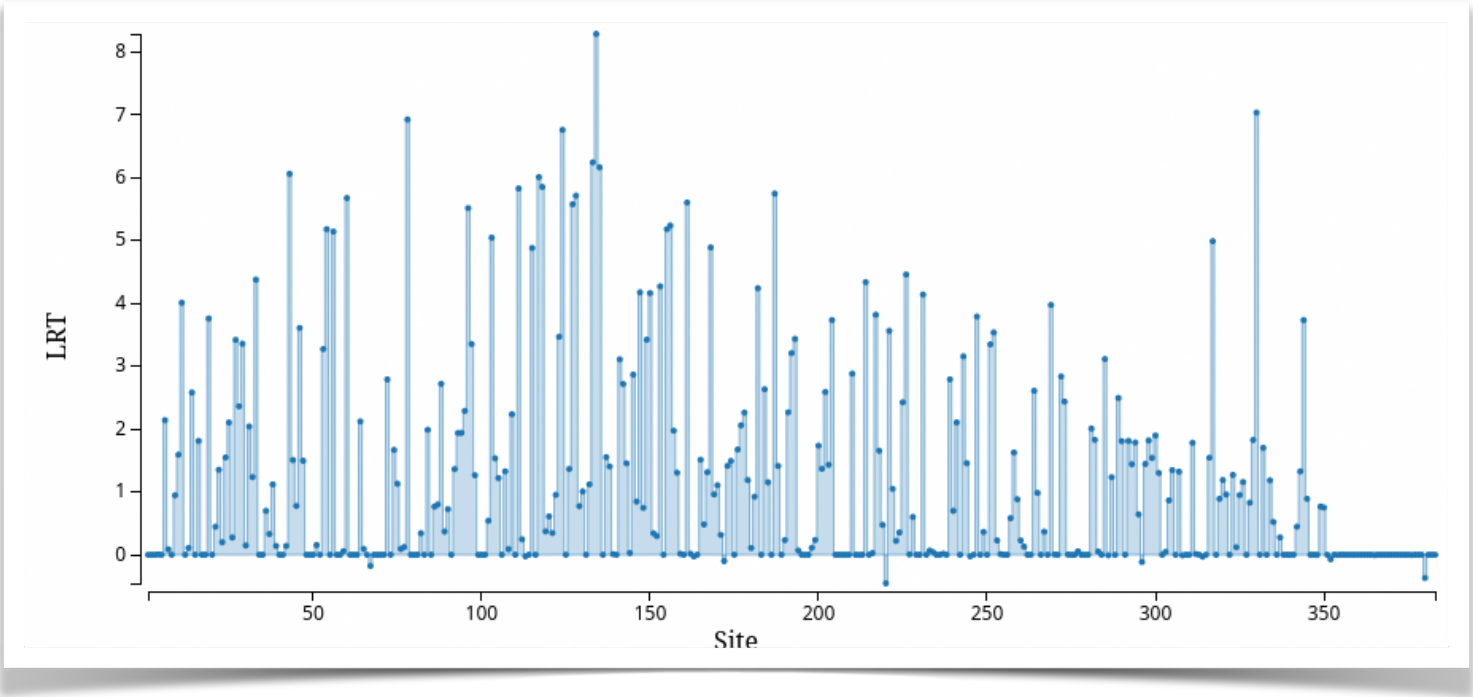


Figure 4. LRT value for each site.

Regarding the use of MEME we checked particularly the sites with positive diversifying selection. The following graph represent the more significant sites within the set and selected by a p-value=0.05 due to the literature of the methodology, selected those positive results and organized it in the following table:

Site	LRT	p-value
134	8.28	
330	7.03	
78	6.92	0.01
124	6.76	
133	6.24	
135	6.16	
43	6.06	
117	6.01	
118	5.85	
111	5.83	0.02
187	5.74	
128	5.71	
60	5.67	
161	5.6	
127	5.57	
96	5.51	
156	5.24	
54	5.18	
155	5.18	0.03
56	5.14	
103	5.04	
317	4.99	
168	4.89	
115	4.88	
226	4.46	
33	4.38	
214	4.34	0.05

Table 1. 25 sites with highest LRT and statistical significance in our 16 paralog OLIG2 gene sample

The 27 sites with the highest LRT were 78, 330, 134, 111, 118, 117, 43, 135, 133, 124, 54, 155, 156, 96, 127, 161, 60, 128, 187, 115, 168, 317, 103, 56, 214, 33, 226. Using the Conserved Domain Database (CDD)(5) and the canonical protein from Uniprot(6) to establish the domain region pertaining to bHLH family in the site 107-168; we determined that about 75%, have of the statistically significant mutations supporting positive diversifying selection were occurring in this domain.

CONCLUSION

Regarding dN/dS we established that indeed there is positive diversifying selection in 27 the OLIG2 gene paralogs and that 75% of sites pertain to the bHLH family. Also established for OLIG2 paralogs an LRT of 4.34 as the cut-off for statistical significance of positive diversification.

3. BLAST: SEARCHING FOR THE HOMOLOGOUS SEQUENCES

A fasta file that contains the query protein sequence of OLIG2_MOUSE was created as *OLIG2.fasta*. The protein database was conducted from Swissprot, from *proteins.fasta* to *proteins.db*.

makeblastdb -in proteins.fasta -dbtype prot -parse_seqids -out proteins.db

blastp -query OLIG2.fasta -db proteins.db -evalue 0.00001 -out **OLIG2protein0.01.out** - new e-value is 0.01

```
Query= sp|Q9EQW6|OLIG2_MOUSE Oligodendrocyte transcription factor 2 OS=Mus
musculus OX=10090 GN=Olig2 PE=1 SV=1

Length=323
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

Q9EQW6 Oligodendrocyte transcription factor 2 OS=Mus musculus OX=...  632     0.0
Q13516 Oligodendrocyte transcription factor 2 OS=Homo sapiens OX=...  610     0.0
Q90XB3 Oligodendrocyte transcription factor 2 OS=Gallus gallus OX...  386     2e-134
Q6PFG8 Oligodendrocyte transcription factor 3 OS=Mus musculus OX=...  257     9e-84
Q7RTU3 Oligodendrocyte transcription factor 3 OS=Homo sapiens OX=...  256     1e-83
Q8BGW3 Class E basic helix-loop-helix protein 23 OS=Mus musculus ...  117     5e-30
Q0V9X5 Class E basic helix-loop-helix protein 22 OS=Xenopus tropi...  117     1e-29
Q8NDY6 Class E basic helix-loop-helix protein 23 OS=Homo sapiens ...  113     9e-29
Q71T09 Class E basic helix-loop-helix protein 22 OS=Gallus gallus...  114     3e-28
Q8NFJ8 Class E basic helix-loop-helix protein 22 OS=Homo sapiens ...  115     4e-28
O09029 Class E basic helix-loop-helix protein 22 OS=Mesocricetus ...  114     7e-28
Q8C6A8 Class E basic helix-loop-helix protein 22 OS=Mus musculus ...  114     7e-28
Q8TAK6 Oligodendrocyte transcription factor 1 OS=Homo sapiens OX=...  98.2    1e-22
O45489 Helix-loop-helix protein 17 OS=Caenorhabditis elegans OX=6...  93.2    2e-22
Q9JKN5 Oligodendrocyte transcription factor 1 OS=Mus musculus OX=...  92.4    1e-20
Q9WUQ3 Oligodendrocyte transcription factor 1 OS=Rattus norvegicu...  92.0    2e-20
P70447 Neurogenin-2 OS=Mus musculus OX=10090 GN=Neurog2 PE=1 SV=1     72.4    2e-13
O09105 Neurogenic differentiation factor 4 OS=Mus musculus OX=100...  68.9    6e-12
P79766 Neurogenic differentiation factor 4 OS=Gallus gallus OX=90...  67.8    2e-11
Q9H2A3 Neurogenin-2 OS=Homo sapiens OX=9606 GN=NEUROG2 PE=1 SV=2      67.0    2e-11
Q9HD90 Neurogenic differentiation factor 4 OS=Homo sapiens OX=960...  67.4    2e-11
Q62414 Neurogenic differentiation factor 2 OS=Mus musculus OX=100...  67.4    2e-11
Q63689 Neurogenic differentiation factor 2 OS=Rattus norvegicus O...  67.4    3e-11
Q15784 Neurogenic differentiation factor 2 OS=Homo sapiens OX=960...  67.4    3e-11
P70562 Class A basic helix-loop-helix protein 15 OS=Rattus norveg...  65.1    3e-11
Q7RTS1 Class A basic helix-loop-helix protein 15 OS=Homo sapiens ...  64.7    3e-11
Q9QYC3 Class A basic helix-loop-helix protein 15 OS=Mus musculus ...  64.7    5e-11
B6VQA1 Protein dimmed OS=Drosophila melanogaster OX=7227 GN=dimm ...  66.6    5e-11
P79765 Neurogenic differentiation factor 1 OS=Gallus gallus OX=90...  64.7    2e-10
Q6NYU3 Neurogenic differentiation factor 6-A OS=Danio rerio OX=79...  64.3    2e-10
Q60867 Neurogenic differentiation factor 1 OS=Mus musculus OX=100...  64.3    2e-10
Q64289 Neurogenic differentiation factor 1 OS=Rattus norvegicus O...  64.3    3e-10
Q13562 Neurogenic differentiation factor 1 OS=Homo sapiens OX=960...  64.3    3e-10
Q60430 Neurogenic differentiation factor 1 OS=Mesocricetus auratu...  64.3    3e-10
Q92886 Neurogenin-1 OS=Homo sapiens OX=9606 GN=NEUROG1 PE=1 SV=2      62.8    4e-10
O42606 Neurogenin-1 OS=Danio rerio OX=7955 GN=neurog1 PE=2 SV=1       61.2    7e-10
P70660 Neurogenin-1 OS=Mus musculus OX=10090 GN=Neurog1 PE=1 SV=1     61.6    8e-10
P70595 Neurogenin-1 OS=Rattus norvegicus OX=10116 GN=Neurog1 PE=2...  61.6    9e-10
```

Figure 5. *Command-line blastp results for OLIG2 homologous*

The top 86 results were extracted from the *OLIG2protein0.01.out* and ID mapping was done in UniProt. ID mapping found 15 ID hits. Their sequences are collected in one fasta file named *blastOLIG2.fasta* which was used as the input for MSA.

clustalo —in=**blastOLIG2.fasta** --out=OLIG2blast.aln --force --outfmt=clustal —wrap=80

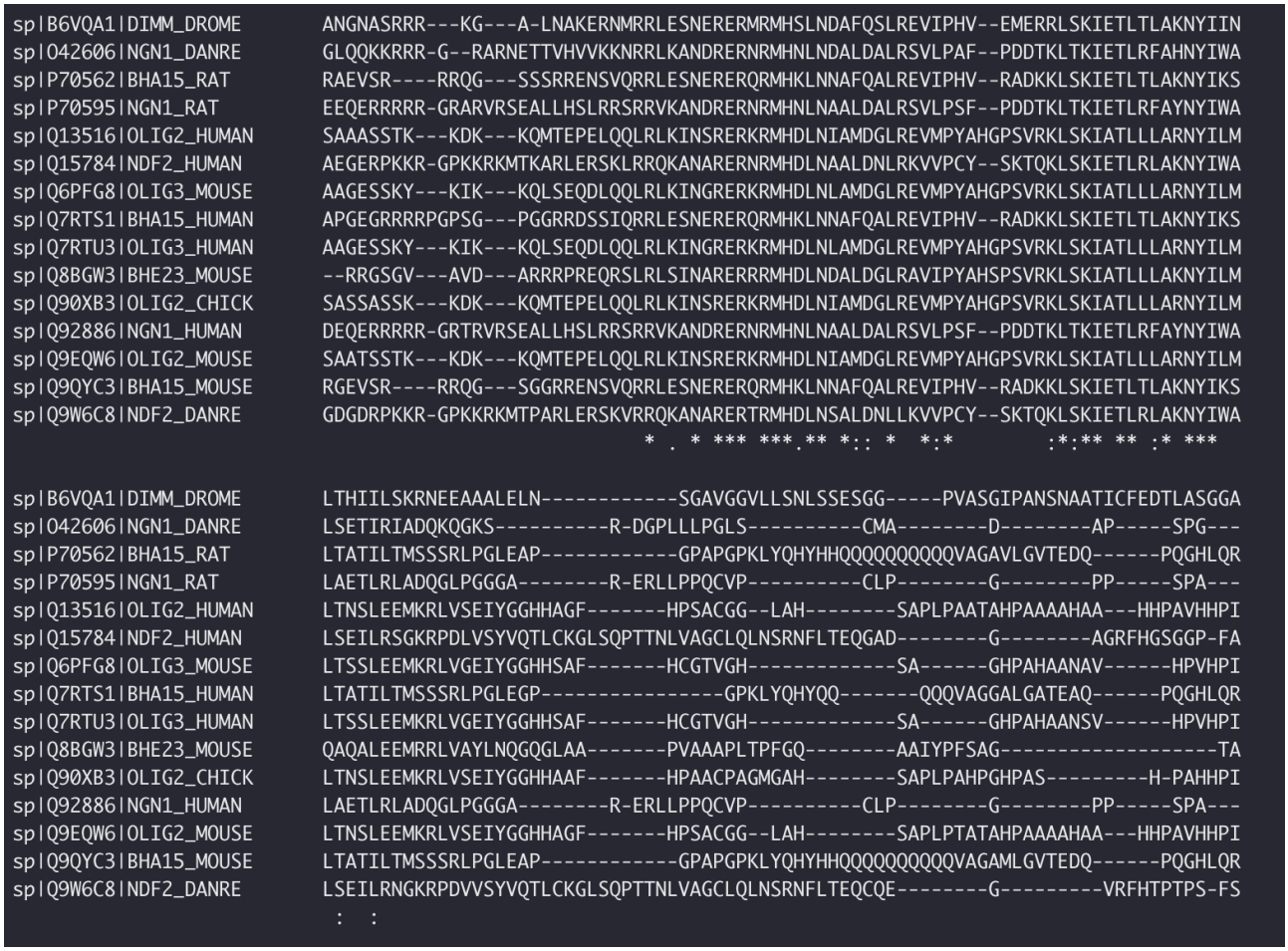


Figure 6. *Command-line OLIG2 homologous MSA*

CD SEARCH

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence. High confidence associations between a query sequence and conserved domains are shown as specific hits.
The query sequence that was used for CD-Search was the protein sequence of OLIG2 in mouse. By preforming CD-Search a conserved region that is mostly located in bHLH superfamily of transcription factors has been found. The output also gave amino acid positions that correspond to this domain. Additionaly, in the output there is a list of homologue proteins that also contain the same conserved domain. As it can be concluded from the picture, most of the homologues are a part off the bHLH supefamily of transcription factors.

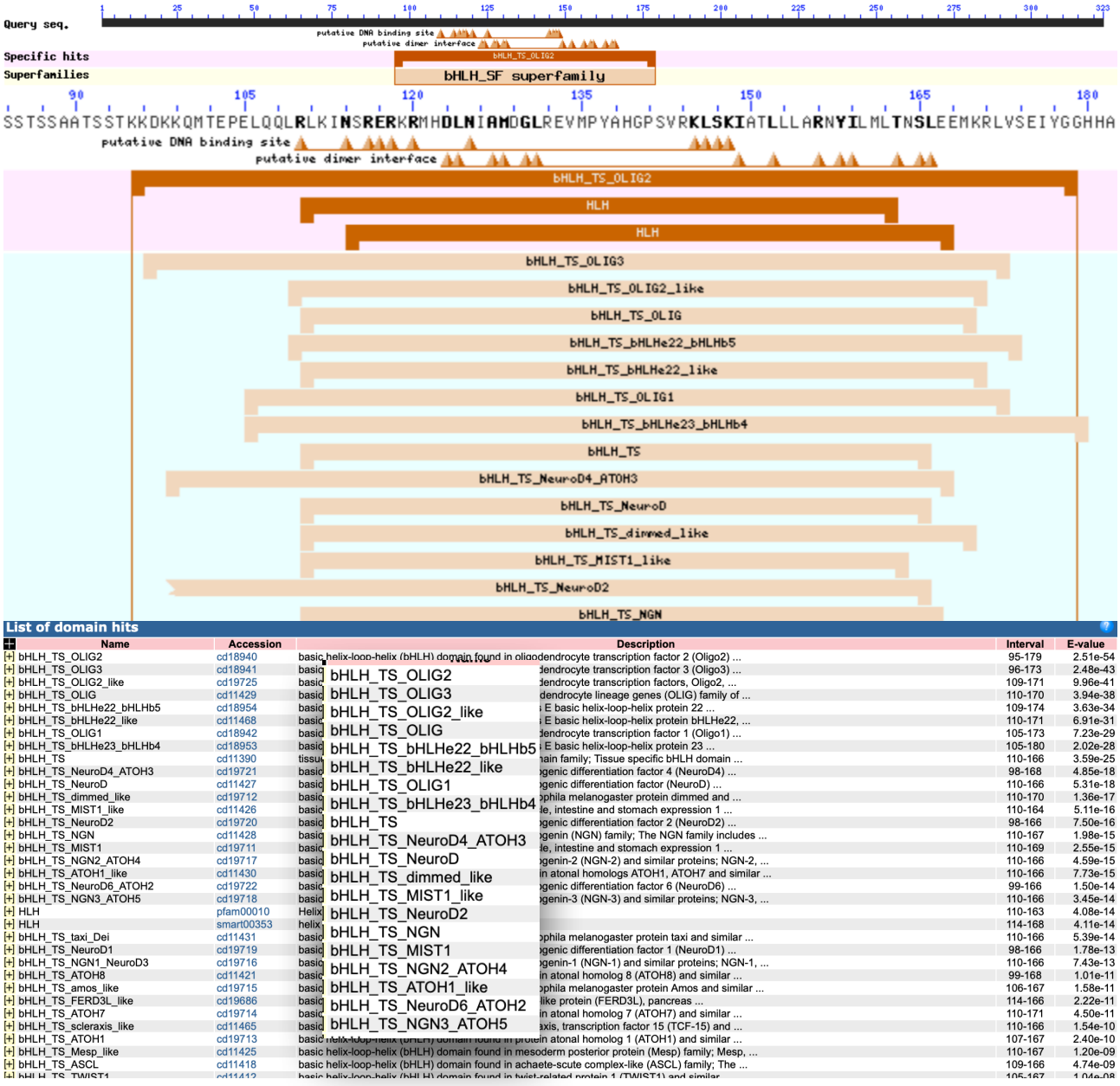


Figure 7. *CD Search OLIG2 domain results showing the conserved bHLH domain and different superfamiliy members*

PSI-BLAST

PSI-BLAST (*Position-Specific Iterative Basic Local Alignment Search Tool*) derives a position-specific scoring matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein–protein BLAST. This PSSM is used to further search the database for new matches, and is updated for subsequent iterations with these newly detected sequences. Thus, PSI-BLAST provides a means of detecting distant relationships between proteins.

PSI-BLAST was preformed with a query protein sequence of OLIG2_MOUSE transcription factor against the protein database conducted from SwissProt. The number of iterations was set to 6, while the E-value was set to 0.005 as this is the standard E-value most commonly used for the best results while iterating through hits.

psiblast -query OLIG2.fasta -db proteins.db -num_iterations=6 -evalue=0.005 -out **OLIG2psiblast.out**

To find out at which line the round 6 begins, used was the corresponding code:

cat -n OLIG2psiblaqst.out | grep "round 6”

To extract only the round 6 results:

sed -n 25333,31795p OLIG2psiblast.out > **OLIG2psiblast6.out** - to extract only the round 6 results

```
→  bin grep "round 6" olig2psiblast6.out
Results from round 6


Query= sp|Q9EQW6|OLIG2_MOUSE Oligodendrocyte transcription factor 2 OS=Mus
musculus OX=10090 GN=Olig2 PE=1 SV=1

Length=323
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value
Sequences used in model and found again:

Q9EQW6 Oligodendrocyte transcription factor 2 OS=Mus musculus OX=...  134     2e-35
Q13516 Oligodendrocyte transcription factor 2 OS=Homo sapiens OX=...  132     8e-35
Q91616 Neurogenic differentiation factor 1 OS=Xenopus laevis OX=8...  119     1e-29
Q91206 Myoblast determination protein 1 homolog 2 OS=Oncorhynchus...  114     2e-28
Q90477 Myoblast determination protein 1 homolog OS=Danio rerio OX...  113     5e-28
O42202 Neurogenic differentiation factor 1 OS=Danio rerio OX=7955...  107     2e-25
A0A3Q7H216 Transcription factor MTB3 OS=Solanum lycopersicum OX=4...  108     6e-25
Q91205 Myoblast determination protein 1 homolog 1 OS=Oncorhynchus...  104     7e-25
Q13562 Neurogenic differentiation factor 1 OS=Homo sapiens OX=960...  104     3e-24
Q8GY61 Transcription factor bHLH63 OS=Arabidopsis thaliana OX=370...  103     5e-24
P21572 Myoblast determination protein 1 homolog OS=Coturnix japon...  99.8    6e-23
Q64289 Neurogenic differentiation factor 1 OS=Rattus norvegicus O...  100     6e-23
P16075 Myoblast determination protein 1 homolog OS=Gallus gallus ...  99.4    9e-23
Q60867 Neurogenic differentiation factor 1 OS=Mus musculus OX=100...  100     9e-23
Q60430 Neurogenic differentiation factor 1 OS=Mesocricetus auratu...  99.8    1e-22
Q15853 Upstream stimulatory factor 2 OS=Homo sapiens OX=9606 GN=U...  99.4    2e-22
Q63665 Upstream stimulatory factor 2 OS=Rattus norvegicus OX=1011...  98.6    3e-22
Q64705 Upstream stimulatory factor 2 OS=Mus musculus OX=10090 GN=...  98.6    3e-22
Q4R5G6 Neurogenic differentiation factor 6 OS=Macaca fascicularis...  97.8    6e-22
Q96NK8 Neurogenic differentiation factor 6 OS=Homo sapiens OX=960...  97.8    6e-22
Q08DI0 Neurogenic differentiation factor 6 OS=Bos taurus OX=9913 ...  97.5    8e-22
O57598 Transcription factor ATOH7 OS=Gallus gallus OX=9031 GN=ATO...  92.8    1e-21
P48986 Neurogenic differentiation factor 6 OS=Mus musculus OX=100...  97.1    1e-21
Q0JXE7 Transcription factor BPE OS=Arabidopsis thaliana OX=3702 G...  95.9    3e-21
Q59RL7 Transcription factor CPH2 OS=Candida albicans (strain SC53...  97.8    3e-21
O49687 Transcription factor MYC4 OS=Arabidopsis thaliana OX=3702 ...  97.1    5e-21
P13904 Myoblast determination protein 1 homolog A OS=Xenopus laev...  94.4    5e-21
P79765 Neurogenic differentiation factor 1 OS=Gallus gallus OX=90...  95.5    6e-21
Q9LNJ5 Transcription factor bHLH13 OS=Arabidopsis thaliana OX=370...  96.3    1e-20
P49811 Myoblast determination protein 1 OS=Sus scrofa OX=9823 GN=...  94.0    1e-20
P20428 Myogenin OS=Rattus norvegicus OX=10116 GN=Myog PE=1 SV=1       92.8    2e-20
Q3T1I5 Sterol regulatory element-binding protein 2 OS=Rattus norv...  95.5    2e-20
```

Figure 8. *Command-line PSI-Blast results for OLIG2 distant homologous*

MSA

The top 170 results were extracted from the *OLIG2psiblast.out* and ID mapping was done in UniProt. ID mapping found 34 ID hits. Their sequences are collected in one fasta file named *OLIG2MSAinput.fasta* which was used as the input for MSA.

clustalo --in=**OLIG2msainput.fasta** --out=OLIG2psiblastmsa.aln --force --outfmt=clustal —wrap=80

```
sp|O49687|MYC4_ARATH          EAESN----RVVVEPEKKP--RKR--------------G---RKPANGR-----EEP-----LNHVEAERQRREKLNQRF
sp|P13904|MYODA_XENLA         EDEHVRAPSGHHQAGRCLL--WAC--------------KACKRKTTN-----------ADRRKAATMRERRRLSKVNEAF
sp|P79765|NDF1_CHICK          -EEEEE--E-EEEDDEQKP--KRR--------------GPKKKKMTKAR-----LERFKLRRMKANARERNRMHGLNAAL
sp|Q9LNJ5|BH013_ARATH         WADAV----GADESGNNRP--RKR--------------G---RRPANGR-----AEA-----LNHVEAERQRREKLNQRF
sp|P49811|MYOD1_PIG           EDEHVRAPSGHHQAGRCLL--WAC--------------KACKRKTTN-----------ADRRKAATMRERRRLSKVNEAF
sp|P20428|MYOG_RAT            EEKGL--GTPEHCPGQCLP--WAC--------------KVCKRKSVS-----------VDRRRAATLREKRRLKKVNEAF
sp|Q3T1I5|SRBP2_RAT           GQEKV------P---------IK----------------QVPGGVKQLE-----PPKEGERRTTHNIIEKRYRSSINDKI
sp|Q07957|USF1_XENBO          SSQDV------LQGGSQRSIAPRT--------------HPYSPKSDGPR-----TTRDDKRRAQHNEVERRRRDKINNWI
sp|P09416|MYC_RAT             -R-------AKLDSGRV----LKQ--------------ISNNRKCSSPR---SSDTEENDKRRTHNVLERQRRNELKRSF
                                                                                          *:     :.  :

sp|O49687|MYC4_ARATH          YSLRAVVPNVS--------------------KMDKASLLGDAISYISELKSKLQKAESDKEELQKQIDVMNKEAGNAKSS
sp|P13904|MYODA_XENLA         ETLKR---YTS---------------TNPNQRLPKVEILRNAIRYIESLQALLHDQD--E-----------------A-
sp|P79765|NDF1_CHICK          DNLRKVVPCYS-----------------KTQKLSKIETLRLAKNYIWALSEILRSGK--SPDL-------------V-S-
sp|Q9LNJ5|BH013_ARATH         YALRSVVPNIS--------------------KMDKASLLGDAVSYINELHAKLKVMEAERERLG----------------
sp|P49811|MYOD1_PIG           ETLKR---CTS---------------SNPNQRLPKVEILRNAIRYIEGLQALLRDQDAAPP-------------GAAAA-
sp|P20428|MYOG_RAT            EALKR---STL---------------LNPNQRLPKVEILRSAIQYIERLQALLSSLNQEERDLRYRG------GGGPSR-
sp|Q3T1I5|SRBP2_RAT           IELKDLVMGTD--------------A-----KMHKSGVLRKAIDYIKYLQQVNHKLRQENMVLKLANQKNKLLKGIDLG-
sp|Q07957|USF1_XENBO          VQLSKIIPDCS--------------MESTKTGQSKGGILSKACDYIQELRQSNLRLSEELQNLDQLQMDNEVLRQQ-VE-
sp|P09416|MYC_RAT             FALRDQIPELE-----------------NNEKAPKVVILKKATAYILSVQADEHKLISEKDLLRKRREQLKHKLEQLRN-
                                *                          *   *     *:  :
```

Figure 9. *Command-line OLIG2 distant homologous MSA output*

Since psiblast finds the most distant homologous, this MSA result represent the region that was evolutionary most conserved. As it is showed, it is not a wide region, compared to most distant homologous sequences.

ERKRMHDLNIAMGLREVMPYAHGPSVRKLSKIATLLLARNYILML (Q9EQW6 | OLIG2_MOUSE)

```
        10         20         30         40         50         60         70         80
MDSDASLVSS RPSSPEPDDL FLPARSKGGS SSGFTGGTVS SSTPSDCPPE LSSELRGAMG ASGAHPGDKL GGGGFKSSSS

        90        100        110        120        130        140        150        160
STSSSTSSAA TSSTKKDKKQ MTEPELQQLR LKINSRERKR MHDLNIAMDG LREVMPYAHG PSVRKLSKIA TLLLARNYIL

       170        180        190        200        210        220        230        240
MLTNSLEEMK RLVSEIYGGH HAGFHPSACG GLAHSAPLPT ATAHPAAAAH AAHHPAVHHP ILPPAAAAAA AAAAAAAVSS

       250        260        270        280        290        300        310        320
ASLPGSGLSS VGSIRPPHGL LKSPSAAAAA PLGGGGGGSG GSGGFQHWGG MPCPCSMCQV PPPHHHVSAM GAGTLPRLTS
```

Figure 10. *UniProt: The positions of amino acids in OLIG2*

Comparing the results to the information provided in the UniProt it is clear that the most conserved region corresponds to a range of amino acids from the position 117 to the position 162 in the peptide which is covered in the bHLH domain (108-162) of the OLIG2 transcription factor. This evolutionary conservation of the bHLH domain points out to its importance in functionality of the OLIG2 transcription factor.

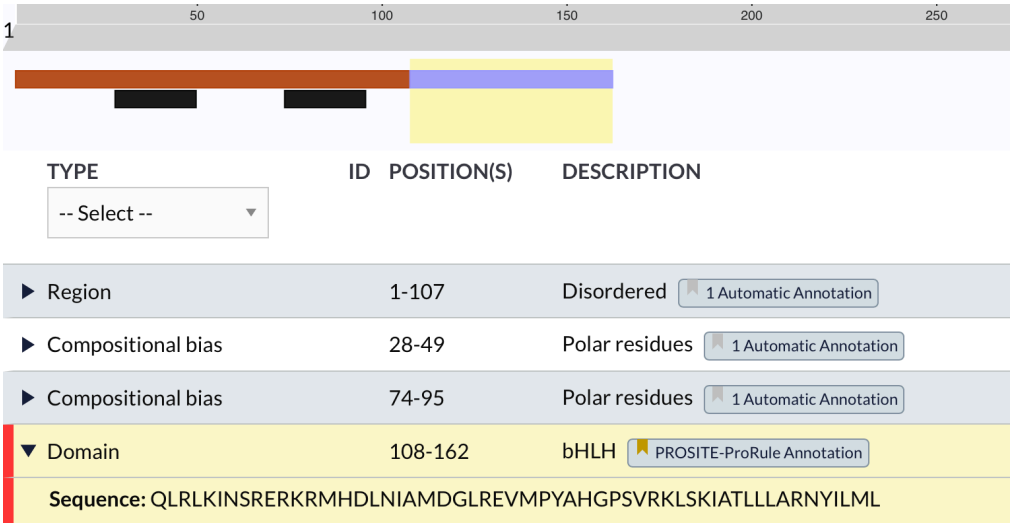


Figure 11. *UniProt bHLH domain position in OLIG2 peptide*

ScanProsite ANALYSIS AND RESULTS

ScanProsite is a web-based tool for detecting signature matches in protein sequences. The tool makes context-dependent annotation templates to detect functional and structural intra-domain residues. In our research ScanProsite was used to check for the DNA-binding domain and the results obtained previously from running the BLAST analysis and CD-search.

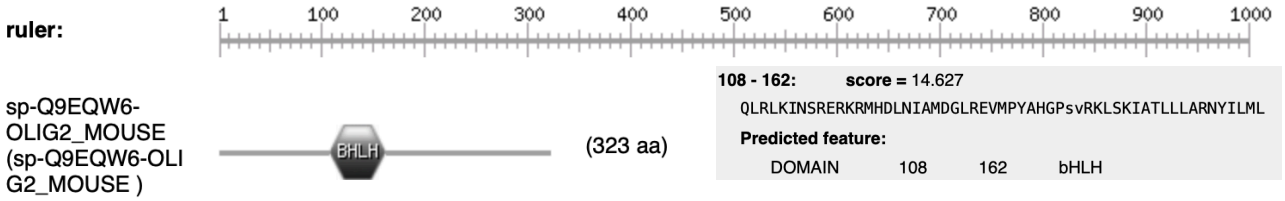


Figure 12. *ScanProsite search results showing the position of bHLH domain of OLIG2*

ScanProsite identified a region approximately between 108 and 162 amino acids' positions and indicated it is a DNA-binding site. By expanding the results a wide description of the bHLH domain is given. This region corresonds to a basic helix-loop-helix domain profile. A number of eukaryotic proteins which are sequence specific DNA-binding proteins that act as transcription factors share a conserved domain of 40 to 50 amino acid residues. It has been proposed that this domain formed of two amphipathic helices joined by a variable length linker region that could form a loop. This domain mediates protein dimerization and has been found in the proteins such as the proteins of the bHLH/PAS superfamily, the myc family of cellular oncogenes, proteins involved in myogenesis, vertebrate neurogenic differentitation factor 1, etc. Members of the bHLH family bind variations on the core sequence 'CANNTG', also refered to as the E-box motif. The homo- or heterodimerization mediated by the HLH domain is independent of, but necessary for DNA binding, as two regions are required for DNA binding activity. The HLH protein lacking the basic domain function as negative regulators since they form heterodimers, but they fail to bind DNA. The cell-type specific members of bHLH superfamily are involved in cell-fate determination and act in neurogenesis, cardiogenesis, myogenesis, and hematopoiesis.

DOMAIN RESEARCH AMONGST THREE OLIG TF IN HUMANS

For the analysis of the conserved domain amongst three OLIG transcription factors in humans used were these three sequences:

>NP_620450.2 OLIG1 [organism=Homo sapiens] [GeneID=116448]
>NP_005797.1 OLIG2 [organism=Homo sapiens] [GeneID=10215]
>NP_786923.1 OLIG3 [organism=Homo sapiens] [GeneID=167826]

clustalo -i olig_all3_human.fasta -o olig_all.aligned
MSA with Jalview > olig_all3_human.aligned.jalview.html

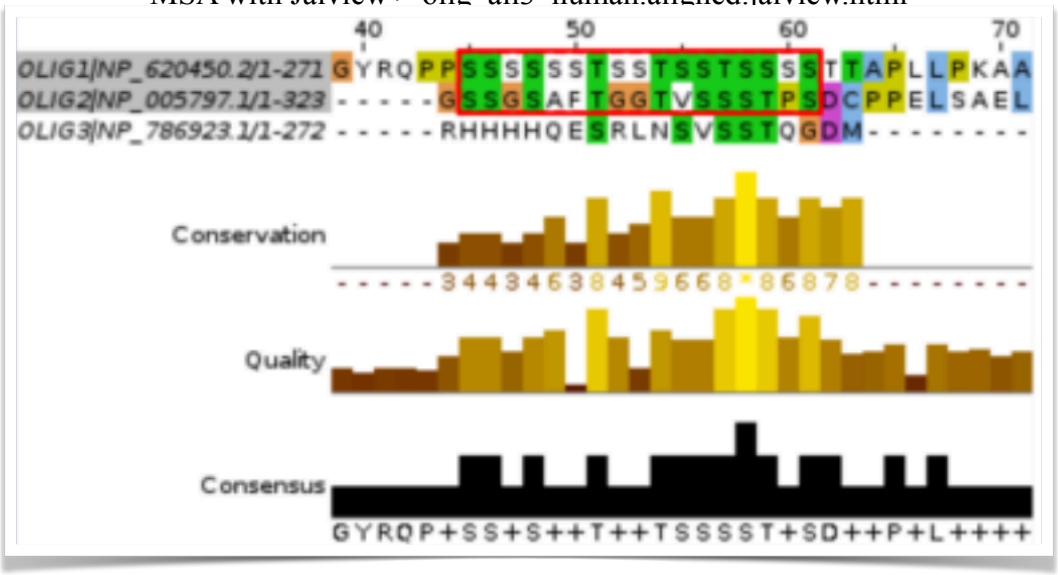


Figure 13. *Domain research results showing conserved positions, quality and consensus amino acids (zoomed in view)*

On the N terminus of the protein there is a one critical region that is shared by OLIG1 and OLIG2 and it is not contained in OLIG3. More importantly, both contain a critical triple serine motif. Notably, the ST box common to OLIG1 and OLIG2 in humans and rodents is not well-conserved in OLIG1 and OLIG2 in other species.

In the middle of the seq, all three contain the bHLH domain. It's worth mentioning that OLIG2 and OLIG3 are more similar than OLIG1. OLIG1 has a longer bHLH domain.

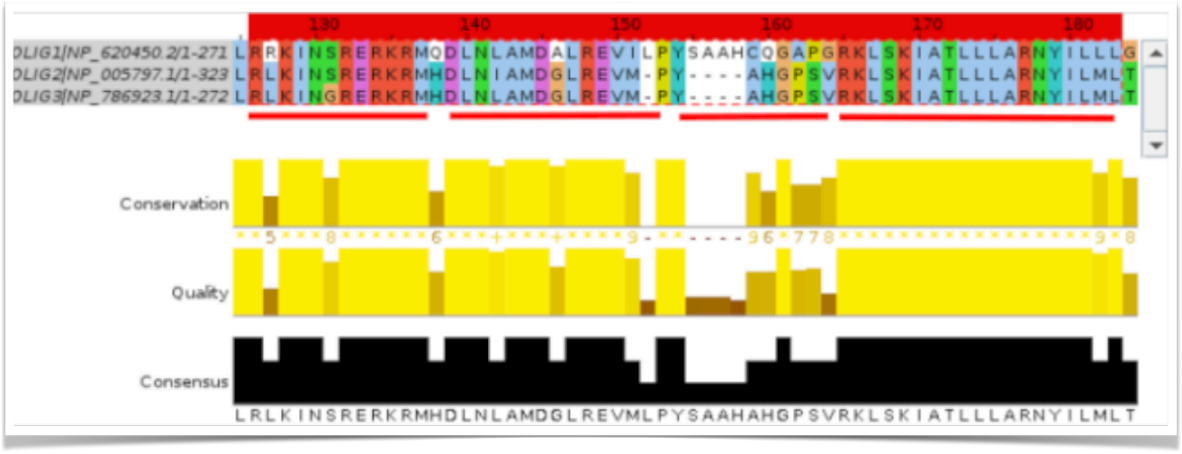
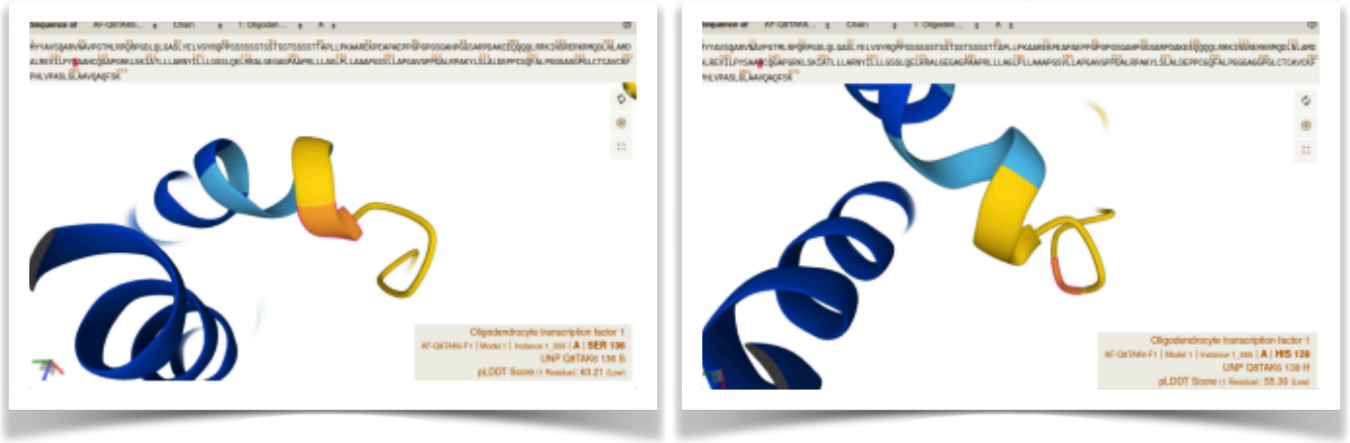


Figure 14. *bHLH domain of three OLIGs*

STRUCTURAL DIFFERENCES BETWEEN OLIG1 AND OLIG2

By analysing and comparing the structure of Olig TFs via *AlphaFold* it is noticable that the sequence -SAAH- in bHLH-domian of OLIG1 is absent in OLIG2 and OLIG3.



https://alphafold.ebi.ac.uk/entry/A0A1A7WSL0

Towards the C-terminus, OLIG2 contains poly-Ala sequence which is not shared with the other two proteins. The only domain close to the C-terminus common to all three OLIG proteins and conserved in all orthologues is a cysteine motif (CXCXXC). Cysteine residues and domains are implicated in multiple roles, including disulphide-bond formation, post-translational modifications such as S−palmitoylation, and recruitment of histone-modifying activities to chromatin.

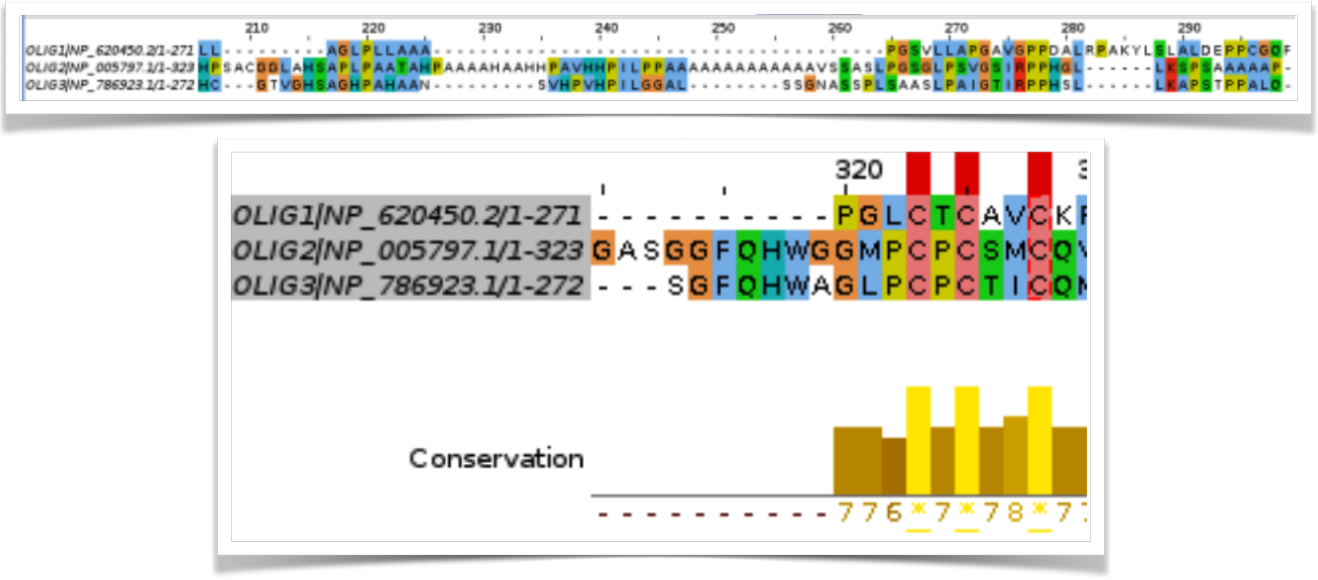


Figure 15. *Comparison of OLIG1, 2 and 3 factors' sequence and their conservation*

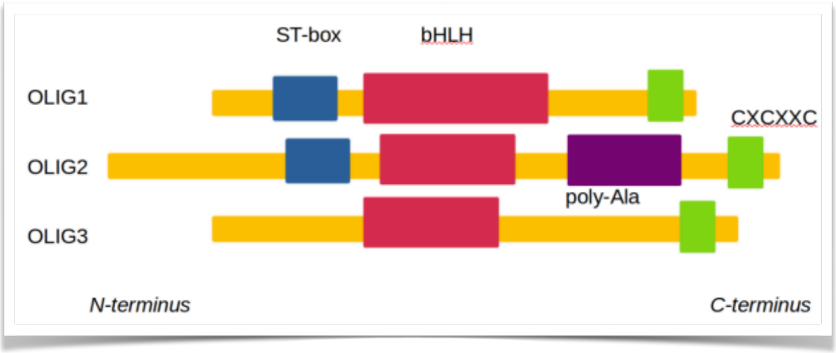


Figure 16. *A schematic view of OLIG1, 2 and 3 domain structure*

PERFORMING MSA ON 7 TF FROM THE SAME FAMILY

clustalo --in=TFbhlh.fasta --out=**output.aln** --force --outfmt=clustal --wrap=80

```
CLUSTAL O(1.2.3) multiple sequence alignment


sp|Q9EQW6|OLIG2_MOUSE      ------------------------------------------------------------------------MDSDASLV
sp|P70447|NGN2_MOUSE       MFVKSETLELKEEEEVLMLLGSASPASATLTPMSSSADEEEDEELRR---PGSAR------------------------
sp|Q62414|NDF2_MOUSE       --------------MLTRLFSEPGLLSDVPKFASWGDGDDDEPRSD---KGDAP------------------------
sp|Q02067|ASCL1_MOUSE      ------------------------------------------------------------------------MESSGKMES
sp|P10085|MYOD1_MOUSE      ------------------------------------------------------------------------------
sp|Q9QX98|PTF1A_MOUSE      ------------MDAVLL--EHFPGGLDTFPS-PYFDEEDFFT--------------------DQSSRDPLEDSDELLG
sp|P48985|ATOH1_MOUSE      -------------------------MSRLLHAEEWAEVKELGDHHRHPQPHHVPPLTPQPPATLQARDLPVYPAELSLL


sp|Q9EQW6|OLIG2_MOUSE      S---------------------------------------SRPSSPE-------PDDLFLPARSKGGSSSGFTGGTVS
sp|P70447|NGN2_MOUSE       --------------GQ---------RGAEAEQ------------------------GVQGSPASGAGGCRPG-------
sp|Q62414|NDF2_MOUSE       ------------------------------------------------------PQPPPAPGSGAPGPARAAKPVS
sp|Q02067|ASCL1_MOUSE      GAGQQ-----PQPPQP---------FLPPAACFFATAAAAAAAAAA----------------------AAQSAQQQQPQA
sp|P10085|MYOD1_MOUSE      -----MELLSPPLRDIDLTGPDGSLCSFETADDFYDDPCFDS--------------PDLRFFED---LDPRLVHMGALLK
sp|Q9QX98|PTF1A_MOUSE      DEQAEVEFLSHQLHEY---------CYRDGACLLLQPAPSAAPHALAPPPLGDPGEPEDNVSYCCDAGAPLAAFPYSPGS
sp|P48985|ATOH1_MOUSE      DSTDPRAWLTPTLQGL---------CTARAAQYLLHSPELGASEAAAPRDE-----ADSQGELVRRSGCGGLSKSPGPVK


sp|Q9EQW6|OLIG2_MOUSE      SST-------PSDCPPELSSELRGAMGASGAHPGDKLGGGGFKSSSSSTSSSTSSAATSSTKK-DK----KQMTEPELQQ
sp|P70447|NGN2_MOUSE       --------------------RL--------------L---GLMHECKRRPS-RSRAVSRGAK---T----AETVQRIKKT
sp|Q62414|NDF2_MOUSE       LRGGEEI-PEPTLAEVKEEGELGGEEEE---EEEEEE---GLDEAEGERP------KKRGPKKRKM----TKARLERSKL
sp|Q02067|ASCL1_MOUSE      PPQQA-----PQLSPVADSQPSGGGHKSAAKQVKRQR------------SSSP--ELMR-CKRRLNFSGFGYSLPQQQPA
sp|P10085|MYOD1_MOUSE      PEEHAHFPTAVHPGPGAREDEH---V---RAPSGHHQAGRCLLWAC------------KACKR-----------KTTNAD
sp|Q9QX98|PTF1A_MOUSE      PPSCLAYPCAAVLSPGA---RLGGLNGAAAAAAAR------------------------RRRRVRSE---AELQ----Q
sp|P48985|ATOH1_MOUSE      VRE--------QLC------KLKGGV------VVDEL---G----CS---------RQ----RAPS----SKQVNGVQKQ


sp|Q9EQW6|OLIG2_MOUSE      LRLKINSRERKRMHDLNIAMDGLREVMPYAHGPSVRKLSKIATLLLARNYILMLTNSLEEMKRLVSEIYGGHHAGFHPSA
sp|P70447|NGN2_MOUSE       RRLKANNRERNRMHNLNAALDALREVLPTFP--EDAKLTKIETLRFAHNYIWALTETLRLADH-----------------
sp|Q62414|NDF2_MOUSE       RRQKANARERNRMHDLNAALDNLRKVVPCYS--KTQKLSKIETLRLAKNYIWALSEILRSGKRPD-------LVSYVQTL
sp|Q02067|ASCL1_MOUSE      AVARRNERERNRVKLVNLGFATLREHVPNGA--ANKKMSKVETLRSAVEYIRALQQLLDEHDAVSAAFQAG---------
sp|P10085|MYOD1_MOUSE      RRKAATMRERRRLSKVNEAFETLKRCTSSN---PNQRLPKVEILRNAIRYIEGLQALLRDQDAAPPGAAAFYAPGPLPPG
sp|Q9QX98|PTF1A_MOUSE      LRQAANVRERRRMQSINDAFEGLRSHIPTLP--YEKRLSKVDTLRLAIGYINFLSELVQADLPLRGSGAGG---------
sp|P48985|ATOH1_MOUSE      RRLAANARERRRMHGLNHAFDQLRNVIPSFN--NDKKLSKYETLQMAQIYINALSELLQTPNVGE-------QPPPPTAS
                                . ***.*:  :* .:  *:           :: *   *  *  **  *   :
```

Figure 17. *Command-line MSA output of 7 transcription factor of the same bHLH superfamily*

According to the results of MSA of the 7 transcription factors of the bHLH familiy a region that is said to be conserved corresponds to the amino acid sequence that ranges from the amino acid position 115 to the position 166 in the OLIG2_MOUSE sequence. This sequence region corresponds to the bHLH domain of this transcription factor (108-162).

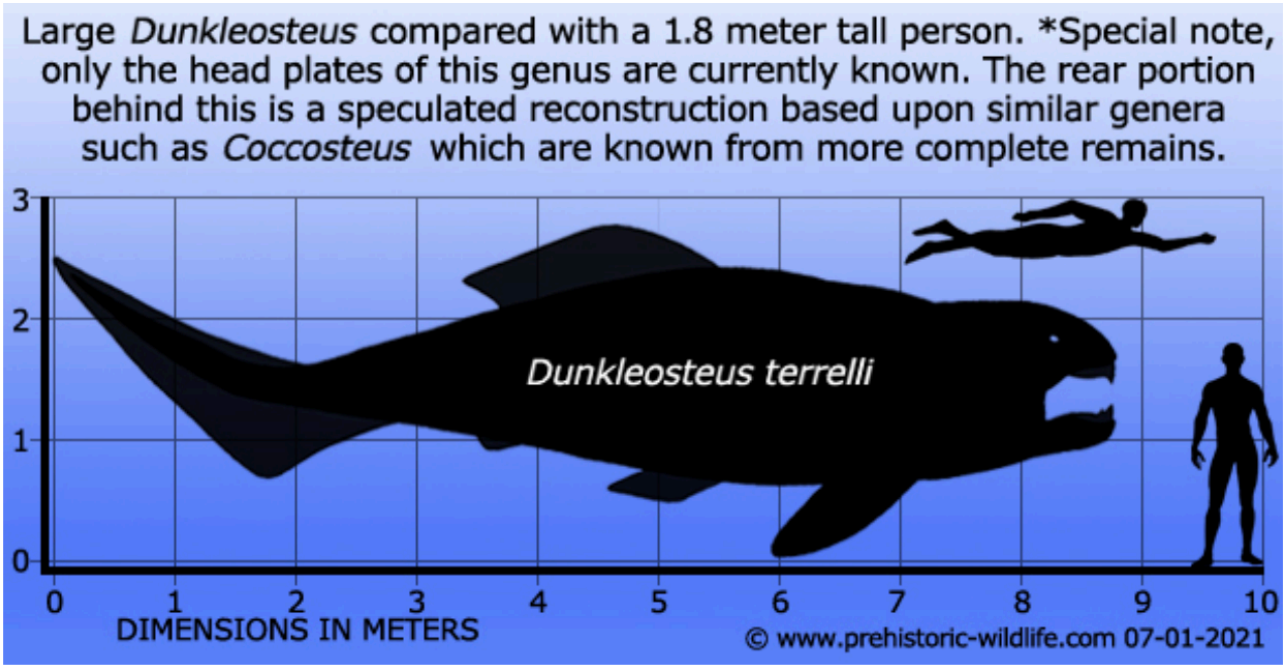
4. PHYLOGENY

IMPORTANCE OF EVOLUTION OF MYELIN

The evolution of myelin enabled axons to propagate electrical impulses at unprecedented speed. This allowed for increase in size, speed as well as ultimately the emergence of intelligence. The cells responsible for myelination are called myelinating cells – oliogodendrocytes in the CNS and Schwann cells in the PNS. Myelin sheaths and myelin-specific proteins are present in Gnathostomata, but are not present in Agnatha.

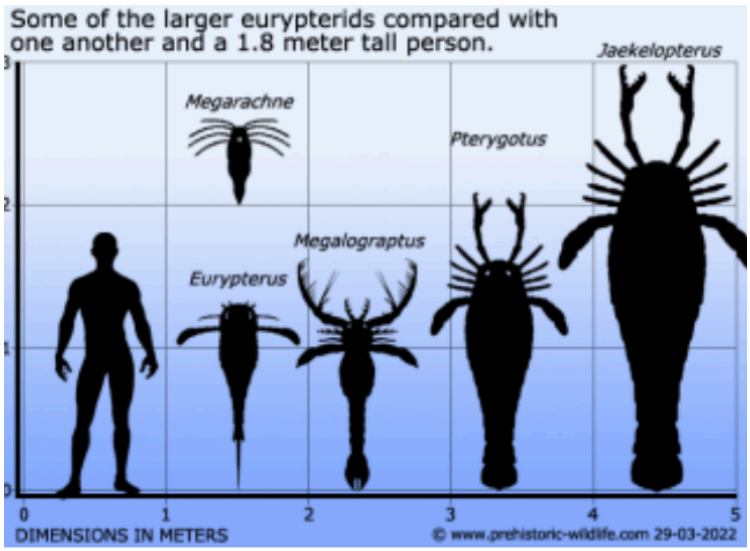
#1 Theory of myelin evolution = evolution of a predatory lifestyle

Placoderms, the earliest jawed fish, were the first vertebrates with myelin (400 Mya) = hinged jaws and myelinated nerves might have evolved in parallel = evolution of a predatory lifestyle.



#2 Theory of myelin evolution = ability to escape from predators

The top predators in Ordovician waters (~450Mya) were arthropods Erypterids or "sea scorpions". These probably preyed on ostracoderms – (jawless) precursors of all present day fish. To escape these predators, myelin evolved. This is supported by the fact that in the spinal cord most oligodendrocytes and motor neurons develop from the same precursor – suggesting that myelin might have evolved to wrap around motor neurons (motor circuits) and to allow for a faster escape reaction.



MSA OF OLIG1, OLIG2 AND OLIG3 IN HUMANS

clustalo -i olig_all3_human.fasta -o olig_all.aligned

MSA was performed on OLIG1, OLIG2 and OLIG3 in humans in order to understand the distance between those three transcriptional factors. The table below represents different scores of the sequences' alignement calculated between OLIG2 and 1, OLIG3 and 1 and OLIG3 and 2, respectivelly. It is evident that the biggest score is between OLIG2 and 3 which is represented below in neighbour joining tree.

Score = 2030.0 Length of alignment = 284 Sequence ***OLIG2***|NP_005797.1/1-275 (Sequence length = 323) Sequence ***OLIG1***|NP_620450.2/12-271 (Sequence length = 271)	***Score = 1910.0*** Length of alignment = 272 Sequence ***OLIG3***|NP_786923.1/1-265 (Sequence length = 272) Sequence ***OLIG1***|NP_620450.2/38-271 (Sequence length = 271)	***Score = 6490.0*** Length of alignment = 323 Sequence ***OLIG3***|NP_786923.1/1-268 (Sequence length = 272) Sequence ***OLIG2***|NP_005797.1/1-323 (Sequence length = 323)

Figure 18. *Tabelar view of the OLIG alignement scores*

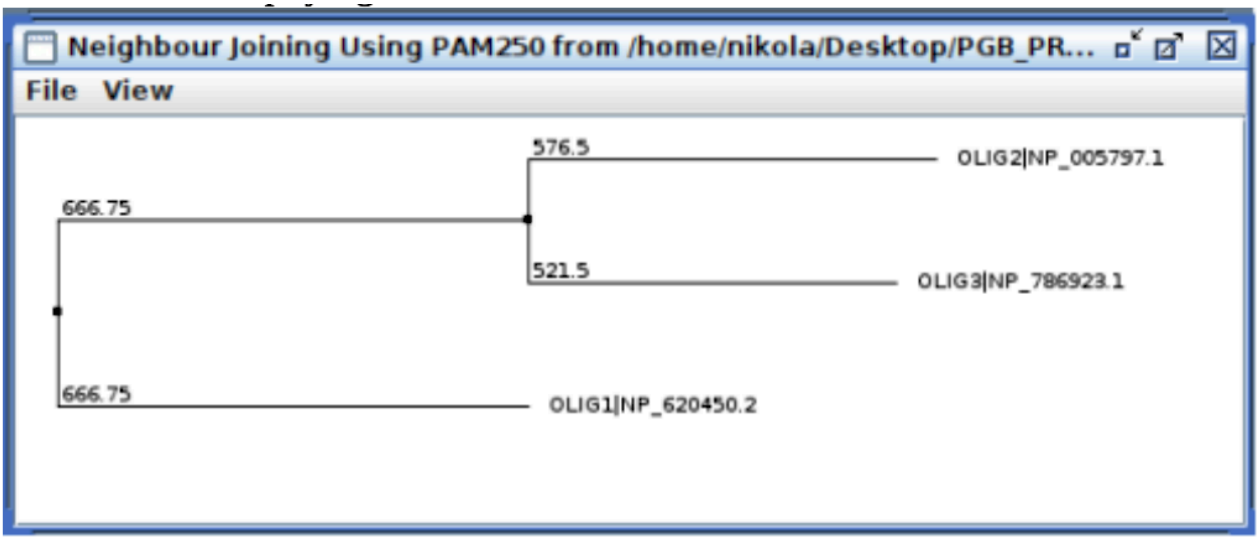


Figure 19. *Neighbour joining tree of OLIGS TF using PAM250 in humans*

MSA OF OLIG1, OLIG2 AND OLIG3 IN DANIO RERIO

To get a sense of the realtion between all four OLIGs. All four are only present in osteichthyes and amphibia. D. rerio was selected because it's wildly used as a model organism, therefore we expect a better annotation of the reference genome.

>NP_001018632.1 oligodendrocyte transcription factor 1 [Danio rerio]
>NP_835201.1 oligodendrocyte transcription factor 2 [Danio rerio]
>NP_001103863.1 oligodendrocyte transcription factor 3 [Danio rerio]
>NP_955808.1 oligodendrocyte transcription factor 4 [Danio rerio]

Observing the neighbour joining tree for OLIGS in Danio rerio it is noticable that OLIG1 is once again more distant from all the other OLIG TF.

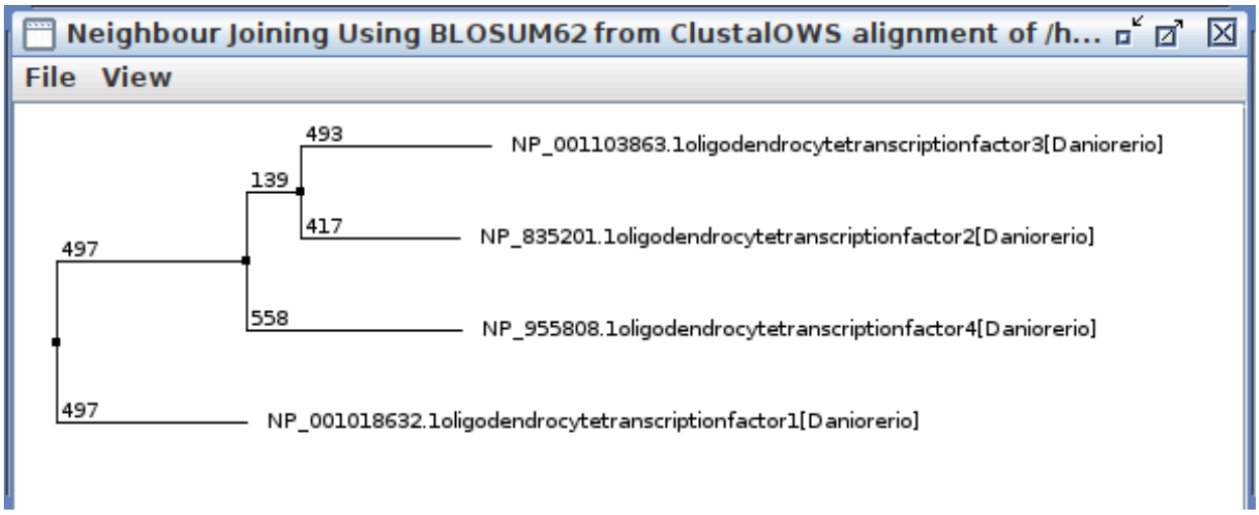


Figure 20. *Neighbour joining tree of OLIGS TF using PAM250 in Dario Rerio*

jalview njt with blosum62 output > olig_all4_drario.aligned.jalview.html

PHYLOGENIC ANALYSIS

A blastp was carried out with the following reference genes: NP_620450.2; NP_005797.1; NP_786923.1; NP_955808.1. The homologues of there genes we searched for in a specified set of species, so that we would have a better understanding of the present olig-genes in different taxa. The set parameters were: expected threshold 0.0005 and scoring matrix blosum80. Blosum80 was used because of the following info from: Bioinformatics - *A Practical Guide to the Analysis of Genes and Proteins (2020) by Andreas D. Baxevanis, B. F. Francis Ouellette.*

The species found along with the different olig-genes present in the species, are specified below:

Species = which olig genes are present
Rattus norvegicus = 1, 2, 3
Mus musculus = 1, 2, 3
Xenopus tropicalis = 1, 2, 3
Anolis carolinensis = 1,
Takifugu rubripes = 1, 2, 3
Danio rerio = 1, 2, 3, 4
Gallus gallus = 2, 3
Saccoglossus kowalevskii = ? (unnamed)
Tetraodon nigroviridis = ? (unnamed)
Callorhinchus milii = 3
Branchiostoma floridae = OligA
Drosophila melanogaster = Olig

OUTPUT: *pylogeny_genes.html*
Sequences were aligned with T-coffee

In the output file *pylogeny_genes.htm* the above mentioned domains are once again apparent.

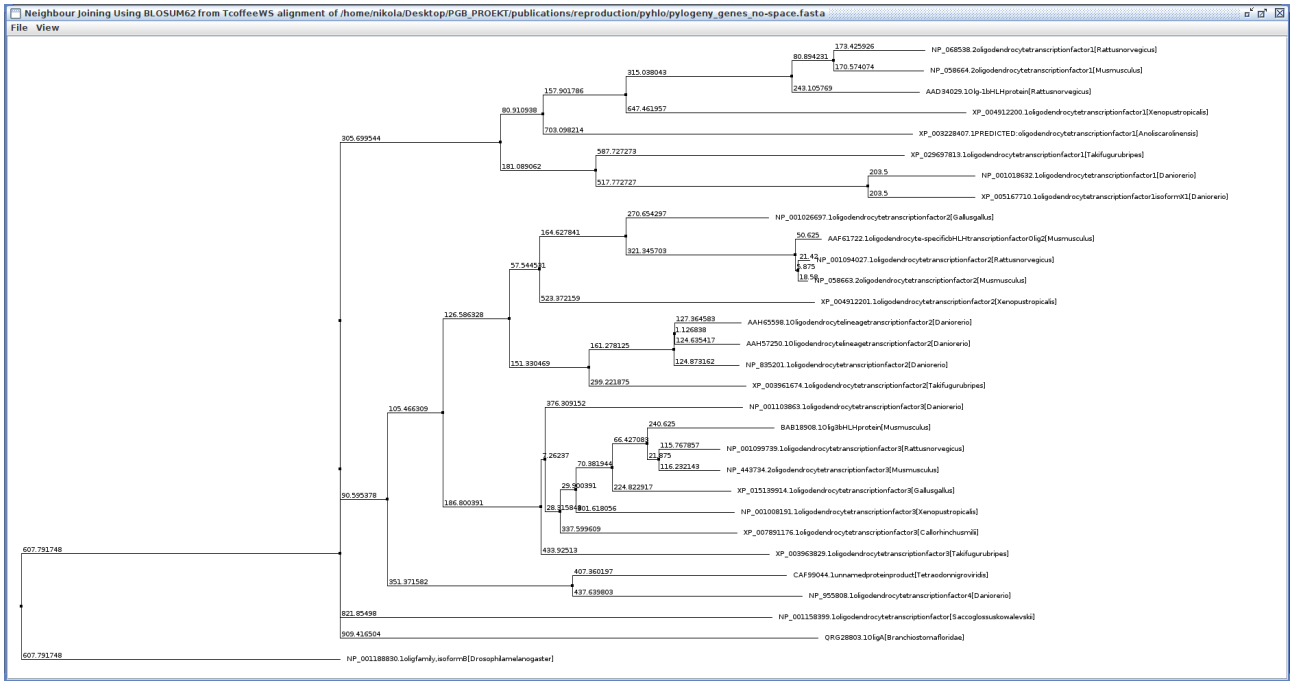


Figure 21. *Neighbour joining tree conducted from different taxa*

With neighbour joining tree the distribution and evoulution of the diffferent OLIG genes is showed. A very interesting pattern that appears is that the different OLIGs are grouped together (OLIG1 genes form one group, OLIG2 another, etc). Furthurmore, it is quite noticable that OLIG2, 3 and 4 are closer to one another, than in respect to OLIG1. A few proto-olig genes have been selected for the analysis (S. kovalevskii, B. floridae, D. melanogaster). They show less distance to OLIG2/3/4 group and more distance to OLIG1 group. This observement gives a basis for the idea that the first functional OLIG was a homologue of OLIG2, 3 or wihich is proposed in the hypothesis below.

LOCATION OF THE GENE

In order to understand the phylogeny the synteny blocks of OLIG genes have been considered.

In humans, the genes encoding oligodendrocyte transcription factor 1 (OLIG1) and OLIG2 are localized within 40 kb of each other on chromosome 21 (syntenic to mouse chromosome 16). Colocalization of these genes is also observed in numerous other species and the chromosomal region in question is well conserved. By contrast, OLIG3 maps to human chromosome 6 (mouse chromosome 10).

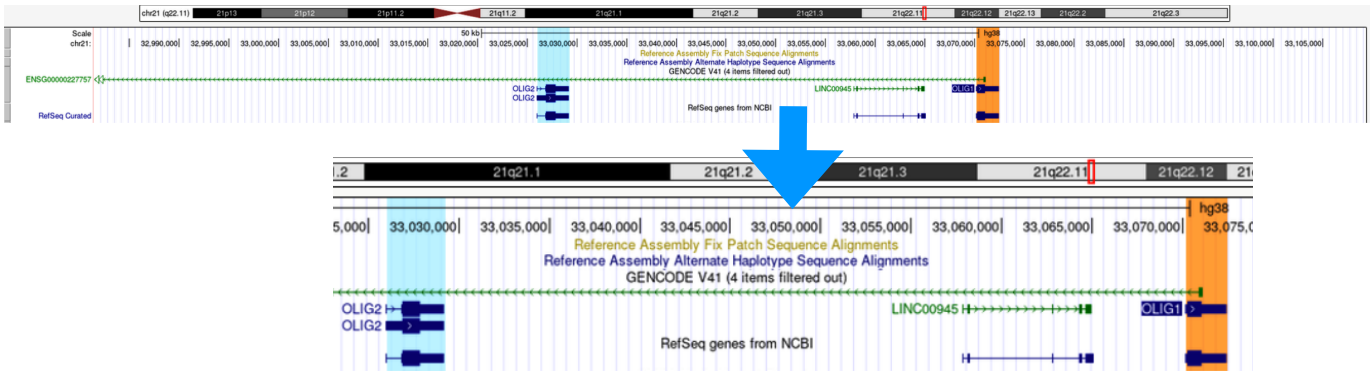


Figure 22. *OLIG2 gene location in humans*

This difference in proximity of genes is a clue for the analysis of phylogeny. In essence, this is a clue that some proto-olig gene was first copied with a whole-genome duplication to give the proto-olig on the other chromosome. Then, with a local genome duplication it would duplicate on same chromosome (forming a synteny block). It is evident that only olig1 and OLIG2 are found in a synthatny block together whereas olig3 is a apart from them. This suggests that this is a proof that will taken into account when doing the filogenetic analysis. On the graph a comparison between human OLIGs and Rattus norvegicus OLIGs are compared.

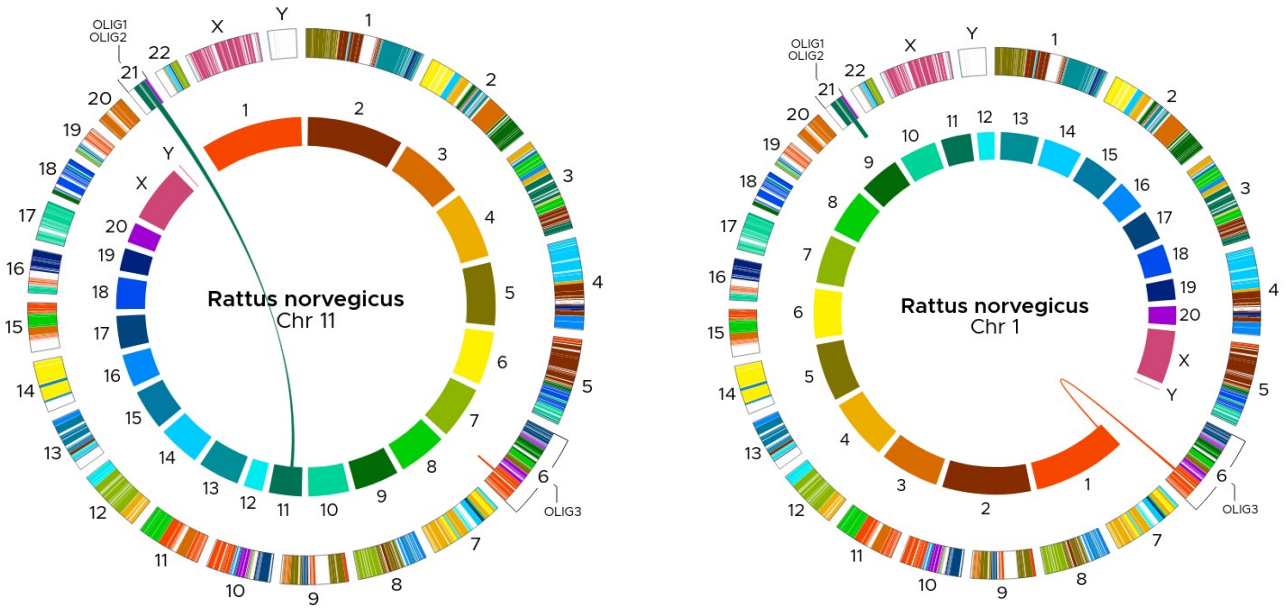
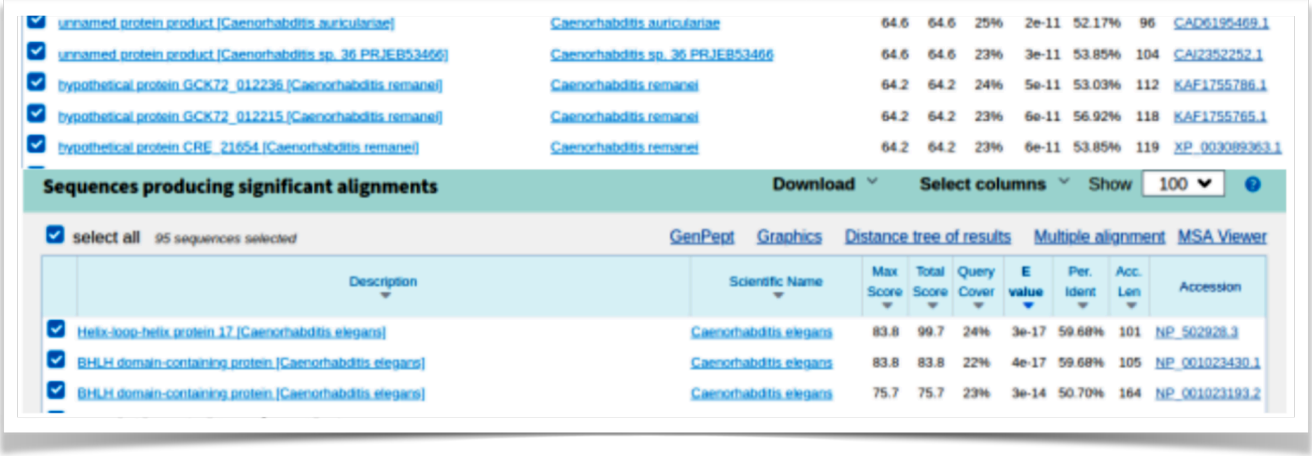


Figure 23. *Synteny blocks between humans and rats. Genes* OLIG1 and OLIG2 are in the same block (A); OLIG3 is in a different block (B). Source: **https://syntenybrowser.jax.org/browse**

INVERTEBRATA

Evidence of a common ancestor for Olig genes can be found in arthropods, nematodes and platyhelminthes. (bHLH homologues can even be found in planst.) In the nematode Caenorhabditis elegans, the Olig homolog Hlh-17 is expressed at all development stages in the cephalic sheath cells (considered to be glial cells) that ensheath four of the dopaminergic neurons and plays a role in dopamine signaling.

Caenorhabditis: (Hlhp17)



unnamed protein product [Caenorhabditis auriculariae]	Caenorhabditis auriculariae	64.6	64.6	25%	2e-11	52.17%	96	CAD6195469.1
unnamed protein product [Caenorhabditis sp. 36 PRJEB53466]	Caenorhabditis sp. 36 PRJEB53466	64.6	64.6	23%	3e-11	53.85%	104	CAI2352252.1
hypothetical protein GCK72_012236 [Caenorhabditis remanei]	Caenorhabditis remanei	64.2	64.2	24%	5e-11	53.03%	112	KAF1755786.1
hypothetical protein GCK72_012215 [Caenorhabditis remanei]	Caenorhabditis remanei	64.2	64.2	23%	6e-11	56.92%	118	KAF1755765.1
hypothetical protein CRE_21654 [Caenorhabditis remanei]	Caenorhabditis remanei	64.2	64.2	23%	6e-11	53.85%	119	XP_003089363.1

Sequences producing significant alignments Download Select columns Show 100

select all 95 sequences selected GenPept Graphics Distance tree of results Multiple alignment MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Helix-loop-helix protein 17 [Caenorhabditis elegans]	Caenorhabditis elegans	83.8	99.7	24%	3e-17	59.68%	101	NP_502928.3
BHLH domain-containing protein [Caenorhabditis elegans]	Caenorhabditis elegans	83.8	83.8	22%	4e-17	59.68%	105	NP_001023430.1
BHLH domain-containing protein [Caenorhabditis elegans]	Caenorhabditis elegans	75.7	75.7	23%	3e-14	50.70%	164	NP_001023193.2

Drosophila: (olig_family)

The Drosophila Olig homolog, Oli, is not expressed in glial lineage cells but in certain ventral motor neuron subtypes; Oli is responsible for regulating larval and adult locomotion and its loss of function can be partially compensated for by over-expression of chick OLIG2.

class E basic helix-loop-helix protein 22 [Drosophila ficusphila]	Drosophila ficusphila	112	133	32%	7e-24	55.21%	232	XP_017044891.1
class E basic helix-loop-helix protein 22 [Drosophila teissieri]	Drosophila teissieri	112	112	29%	7e-24	55.21%	232	XP_043643965.1
class E basic helix-loop-helix protein 22 [Drosophila rhopaloa]	Drosophila rhopaloa	112	133	32%	7e-24	55.21%	232	XP_016975134.1
LOW QUALITY PROTEIN: class E basic helix-loop-helix protein 22 [Drosophila elegans]	Drosophila elegans	112	133	32%	7e-24	55.21%	232	XP_017117172.1
olig family, isoform B [Drosophila melanogaster]	Drosophila melanogaster	112	112	29%	7e-24	55.21%	232	NP_001188830.1
class E basic helix-loop-helix protein 22 [Drosophila erecta]	Drosophila erecta	112	112	29%	7e-24	55.21%	232	XP_001974428.1
class E basic helix-loop-helix protein 22 [Drosophila mauritiana]	Drosophila mauritiana	112	112	29%	7e-24	55.21%	232	XP_033173486.1
class E basic helix-loop-helix protein 22 [Drosophila suzukii]	Drosophila suzukii	112	133	32%	7e-24	55.21%	233	XP_016945403.1
class E basic helix-loop-helix protein 22 [Drosophila biarmipes]	Drosophila biarmipes	112	133	32%	7e-24	55.21%	233	XP_016964404.1

Agnatha: (oligA and OLIG2-like)

As jawless fish (agnatha) are the only vertebrates without compact myelin, they are important phylogenetic tools for studying Olig gene evolution. Cartilaginous fish (family Chondrichthyes) are the most ancient living species to possess myelin.

oligodendrocyte transcription factor 2-like [Petromyzon marinus]	Petromyzon marinus	90.9	90.9	26%	1e-20	70.42%	363	XP_032832704.1
uncharacterized protein LOC116953313 [Petromyzon marinus]	Petromyzon marinus	89.0	89.0	24%	7e-20	73.13%	384	XP_032829299.1
OligA [Lampetra planeri]	Lampetra planeri	77.8	77.8	22%	4e-18	75.41%	86	AZK65268.1
class E basic helix-loop-helix protein 22-like [Petromyzon marinus]	Petromyzon marinus	72.0	72.0	26%	6e-14	58.33%	381	XP_032809050.1
neurogenic differentiation factor 1 [Petromyzon marinus]	Petromyzon marinus	56.2	56.2	26%	1e-08	45.83%	431	XP_032802247.1
neurogenic differentiation factor [Lethenteron camtschaticum]	Lethenteron camtschaticum	50.8	50.8	23%	4e-08	49.21%	100	BBG56411.1
neurogenin [Lethenteron camtschaticum]	Lethenteron camtschaticum	52.4	52.4	24%	1e-07	46.38%	256	BBG56412.1
neurogenin-2 [Petromyzon marinus]	Petromyzon marinus	52.0	52.0	24%	2e-07	46.38%	280	XP_032825193.1
neurogenic differentiation factor 2-like isoform X2 [Petromyzon marinus]	Petromyzon marinus	52.4	52.4	26%	2e-07	45.83%	325	XP_032821678.1
neurogenic differentiation factor 1-like [Petromyzon marinus]	Petromyzon marinus	52.4	52.4	22%	3e-07	49.18%	524	XP_032807768.1
protein dimmed-like isoform X1 [Petromyzon marinus]	Petromyzon marinus	52.0	52.0	26%	3e-07	45.83%	440	XP_032821677.1
class A basic helix-loop-helix protein 15-like [Petromyzon marinus]	Petromyzon marinus	47.4	47.4	25%	6e-06	47.83%	262	XP_032834083.1
mastermind-like domain-containing protein 1 [Petromyzon marinus]	Petromyzon marinus	41.6	41.6	22%	5e-04	38.46%	324	XP_032800822.1
Neurogenin A [Lampetra planeri]	Lampetra planeri	36.6	36.6	11%	0.016	56.67%	201	CEP28078.1

MSA was carried out to understand the relation of proto-oligs to human OLIGs. Sequences were aligned with T-coffee and the neighbour joining tree was constructed with PAM250. Sequences that have been used were:

NP_502928.3
AZK65268.1
XP_032832704.1
NP_001188830.1
NP_620450.2
NP_005797.1
NP_786923.1

(T-coffee PAM250) → output in *agnatha_aligned.jalview.html*

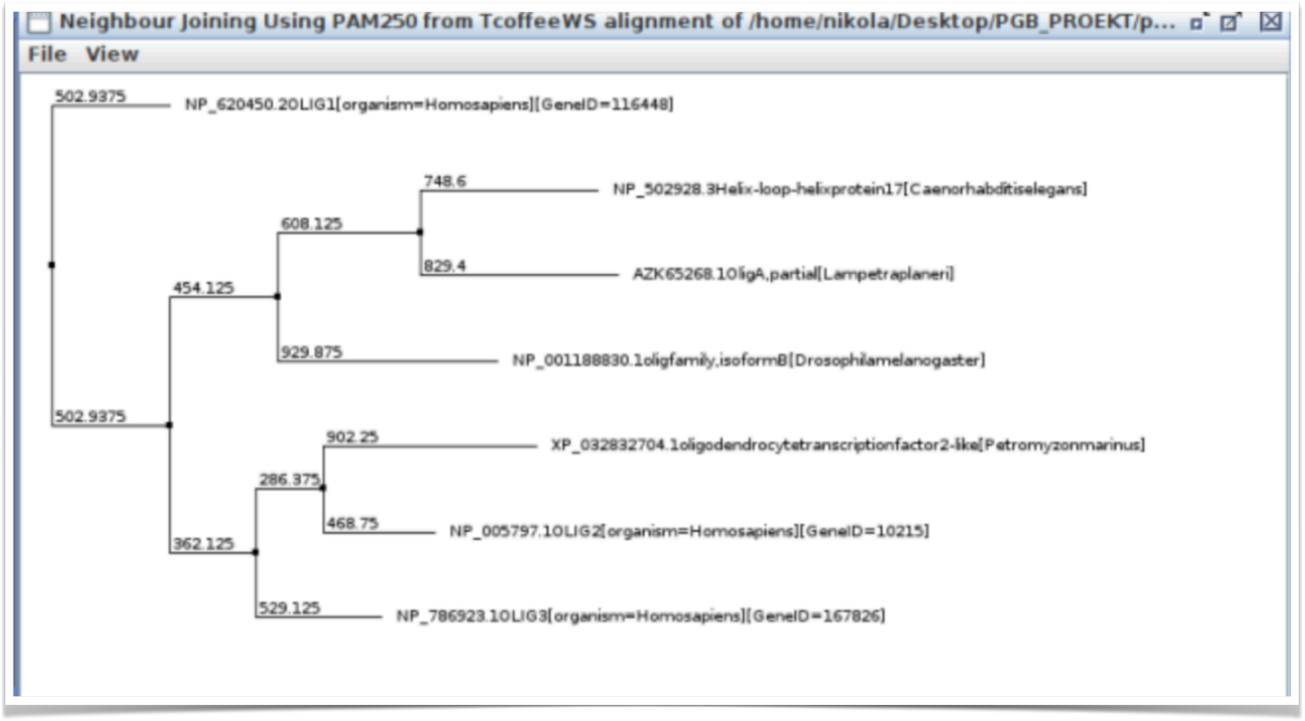


Figure 24. *Neighbour joining tree of selected Agnatha species and humans, using PAM250*

It is showed that OLIG1 is once more further away from proto-oligs giving support for our hypothesis.

VERTEBRATA

It is interesting to note that the elephant shark (Chondrichthyes) has two OLIG2 homologues just about 20 kb apart in one scaffold, which is similar to the way that Olig1 and OLIG2 are located in a synteny block in bony vertebrates, lending support to our previous hypothesis that Olig1 might have evolved from a local duplication at an ancestral OLIG2

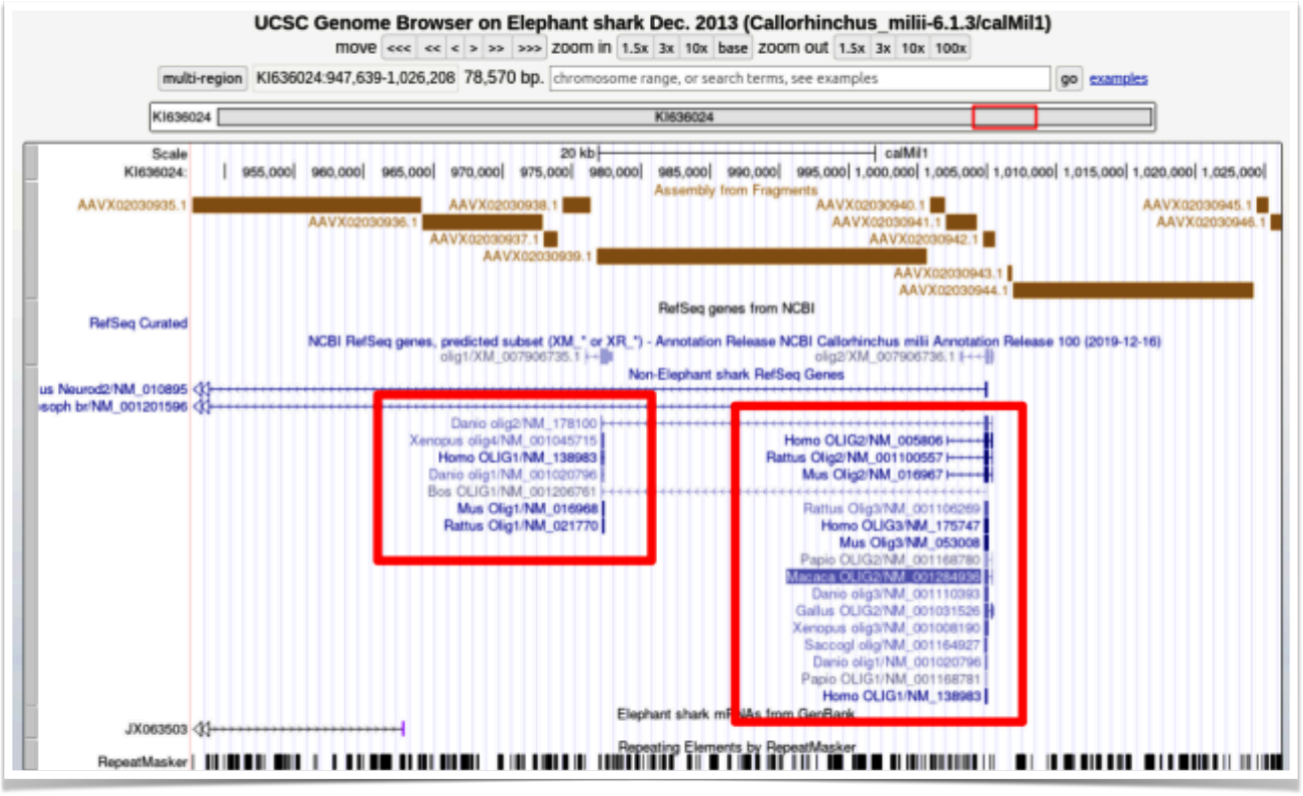


Figure 25. *USCS Genome Browser on Elephant shark*

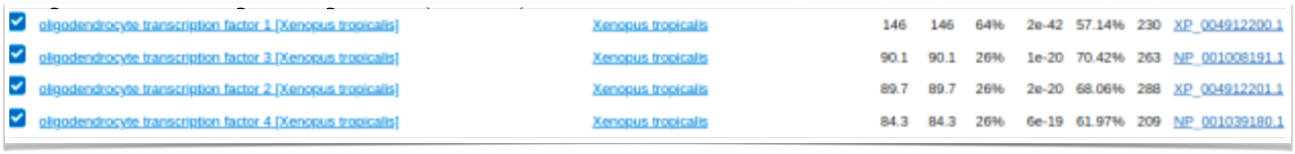
Chondrichthyes = Callorhinchus milii (1,2,3)

oligodendrocyte transcription factor 3-like [Scyliorhinus canicula]	Scyliorhinus canicula	100	100	64%	3e-24	44.94%	216	XP_038657585.1
oligodendrocyte transcription factor 3-like [Carcharodon carcharias]	Carcharodon carcharias	99.8	99.8	64%	4e-24	44.38%	216	XP_041066796.1
oligodendrocyte transcription factor 3-like [Chiloscyllium plagiosum]	Chiloscyllium plagiosum	99.4	99.4	64%	5e-24	43.58%	217	XP_043557250.1
hypothetical protein [Chiloscyllium punctatum]	Chiloscyllium punctatum	99.4	99.4	64%	6e-24	43.58%	217	GCC30474.1
oligodendrocyte transcription factor 1 [Amblyraja radiata]	Amblyraja radiata	95.9	95.9	64%	9e-23	42.70%	216	XP_032888460.1
oligodendrocyte transcription factor 3-like [Rhincodon typus]	Rhincodon typus	95.9	95.9	64%	1e-22	42.46%	217	XP_020367395.1
oligodendrocyte transcription factor 3-like [Stegostoma fasciatum]	Stegostoma fasciatum	95.5	95.5	64%	1e-22	42.46%	217	XP_048396166.1
oligodendrocyte transcription factor 2-like [Rhincodon typus]	Rhincodon typus	91.7	91.7	26%	7e-21	69.44%	238	XP_020367394.1
oligodendrocyte transcription factor 2-like [Chiloscyllium plagiosum]	Chiloscyllium plagiosum	91.7	91.7	26%	7e-21	69.44%	241	XP_043557533.1

Osteichthyes = Danio rerio (1,2,3,4)

oligodendrocyte transcription factor 1 [Danio rerio]	Danio rerio	109	109	27%	2e-25	64.77%	235	NP_001018632.1
oligodendrocyte transcription factor 1 isoform X1 [Danio rerio]	Danio rerio	109	109	27%	2e-25	64.77%	235	XP_005167710.1
oligodendrocyte transcription factor 3 [Danio rerio]	Danio rerio	94.4	94.4	26%	2e-20	70.42%	255	NP_001103863.1
Oligodendrocyte lineage transcription factor 2 [Danio rerio]	Danio rerio	93.1	93.1	26%	8e-20	66.20%	273	AAH65598.1
Oligodendrocyte lineage transcription factor 2 [Danio rerio]	Danio rerio	93.1	93.1	26%	8e-20	66.20%	273	AAH57250.1
oligodendrocyte transcription factor 2 [Danio rerio]	Danio rerio	93.1	93.1	26%	8e-20	66.20%	273	NP_835201.1
oligodendrocyte transcription factor 4 [Danio rerio]	Danio rerio	91.8	91.8	27%	2e-19	64.86%	244	NP_955808.1
class E basic helix-loop-helix protein 23 [Danio rerio]	Danio rerio	75.4	75.4	26%	3e-14	58.33%	227	NP_001025304.1

Amphibia = Xenopus tropicalis (1,2,3,4)



oligodendrocyte transcription factor 1 [Xenopus tropicalis]	Xenopus tropicalis	146	146	64%	2e-42	57.14%	230	XP_004912200.1
oligodendrocyte transcription factor 3 [Xenopus tropicalis]	Xenopus tropicalis	90.1	90.1	26%	1e-20	70.42%	263	NP_001008191.1
oligodendrocyte transcription factor 2 [Xenopus tropicalis]	Xenopus tropicalis	89.7	89.7	26%	2e-20	68.06%	288	XP_004912201.1
oligodendrocyte transcription factor 4 [Xenopus tropicalis]	Xenopus tropicalis	84.3	84.3	26%	6e-19	61.97%	209	NP_001039180.1

Reptilia (1,2,3)

oligodendrocyte transcription factor 1 [Podarcis muralis]	Podarcis muralis	133	190	47%	1e-36	93.24%	234	XP_028582888.1
oligodendrocyte transcription factor 1 [Zootoca vivipara]	Zootoca vivipara	133	194	47%	2e-36	93.24%	238	XP_034972055.1
oligodendrocyte transcription factor 1 [Sceloporus undulatus]	Sceloporus undulatus	132	174	36%	3e-36	92.96%	242	XP_042315529.1
PREDICTED: oligodendrocyte transcription factor 1 [Thamnophis sirtalis]	Thamnophis sirtalis	132	197	48%	6e-36	91.55%	231	XP_013923546.1
PREDICTED: oligodendrocyte transcription factor 1-like [Gekko japonicus]	Gekko japonicus	124	124	62%	2e-33	58.38%	204	XP_015268977.1
oligodendrocyte transcription factor 2-like [Pseudonaja textilis]	Pseudonaja textilis	89.7	89.7	26%	2e-20	71.83%	183	XP_026581991.1
oligodendrocyte transcription factor 2 [Python bivittatus]	Python bivittatus	91.3	91.3	26%	3e-20	69.01%	256	XP_007434497.1
oligodendrocyte transcription factor 3 [Pogona vitticeps]	Pogona vitticeps	89.4	89.4	26%	3e-20	73.24%	189	XP_020655954.1
oligodendrocyte transcription factor 3 [Python bivittatus]	Python bivittatus	90.1	90.1	26%	4e-20	70.42%	218	XP_007429853.1
oligodendrocyte transcription factor 2 [Pogona vitticeps]	Pogona vitticeps	89.7	89.7	26%	5e-20	69.01%	221	XP_020651891.1
oligodendrocyte transcription factor 2 [Thamnophis elegans]	Thamnophis elegans	91.3	91.3	26%	6e-20	69.01%	308	XP_032075069.1
PREDICTED: oligodendrocyte transcription factor 2 [Thamnophis sirtalis]	Thamnophis sirtalis	91.3	91.3	26%	6e-20	69.01%	308	XP_013923535.1

Aves (2, 3)

oligodendrocyte transcription factor 2 [Meleagris gallopavo]	Meleagris gallopavo	89.4	89.4	26%	4e-21	71.83%	138	XP_010725796.1
oligodendrocyte transcription factor 2 [Centrocercus urophasianus]	Centrocercus urophasianus	91.3	91.3	26%	2e-20	69.01%	300	XP_042666798.1
oligodendrocyte transcription factor 2 [Phasianus colchicus]	Phasianus colchicus	90.9	90.9	26%	3e-20	69.01%	299	XP_031468509.1
hypothetical protein CIB84_009387 [Bambusicola thoracicus]	Bambusicola thoracicus	90.9	90.9	26%	3e-20	69.01%	298	POI26863.1
oligodendrocyte transcription factor 2 [Aquila chrysaetos chrysaetos]	Aquila chrysaetos chrysaetos	90.5	90.5	26%	4e-20	69.01%	291	XP_029876914.1
LOW QUALITY PROTEIN: oligodendrocyte transcription factor 3 [Aquila chrysaetos chrysaetos]	Aquila chrysaetos chrysaetos	89.7	89.7	26%	6e-20	70.42%	273	XP_029878819.1
hypothetical protein CIB84_015656 [Bambusicola thoracicus]	Bambusicola thoracicus	89.7	89.7	26%	6e-20	70.42%	273	POI20597.1
LOW QUALITY PROTEIN: oligodendrocyte transcription factor 2 [Coturnix japonica]	Coturnix japonica	90.1	90.1	26%	6e-20	67.61%	301	XP_015739081.1
oligodendrocyte transcription factor 3 [Meleagris gallopavo]	Meleagris gallopavo	89.7	89.7	26%	6e-20	70.42%	273	XP_010705555.1
oligodendrocyte transcription factor 3 [Coturnix japonica]	Coturnix japonica	89.0	89.0	26%	9e-20	69.01%	272	XP_032299766.1

In mammalians all three OLIGs are present.

CONCLUSIONS

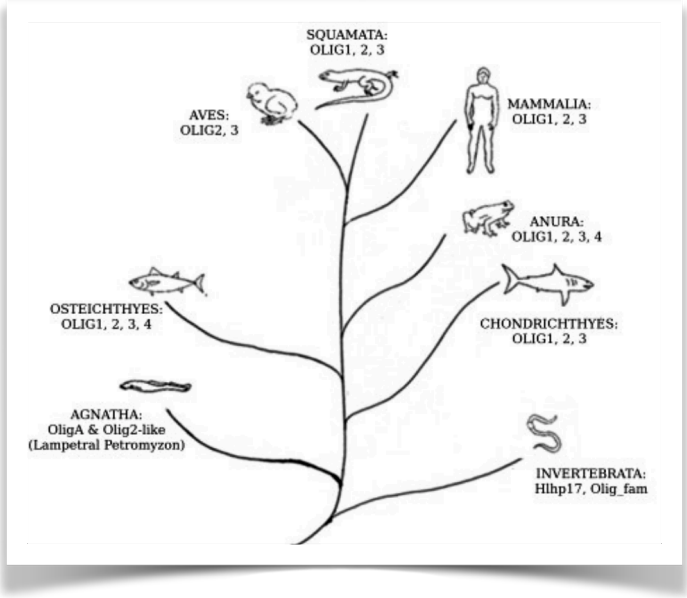


Figure 26. *Schematic representing the found homologues*

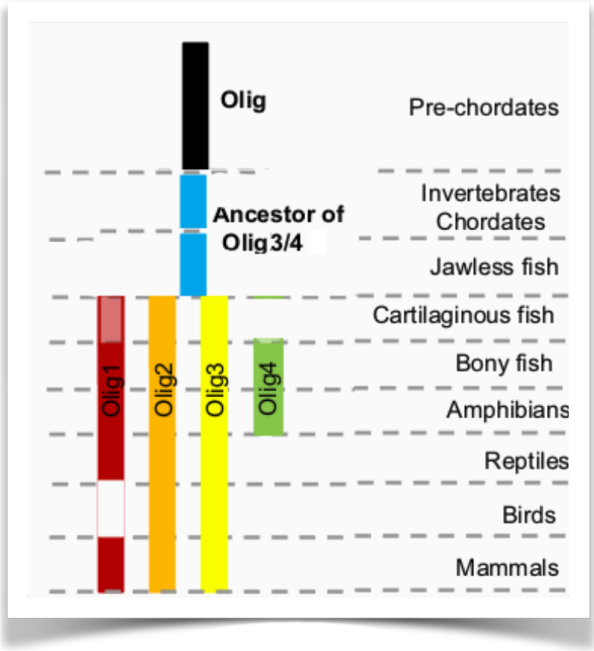


Figure 27. *Diagram the found homologues and their annotation in different taxa*

Both schemes hint that the homologues in Agnatha are the precursors to OLIG genes in the Gnathostomata. Taken into consideration that the homologues in Agnatha are closer to the OLIG2 and OLIG3 genes than the rest, the following can be hypothesized:

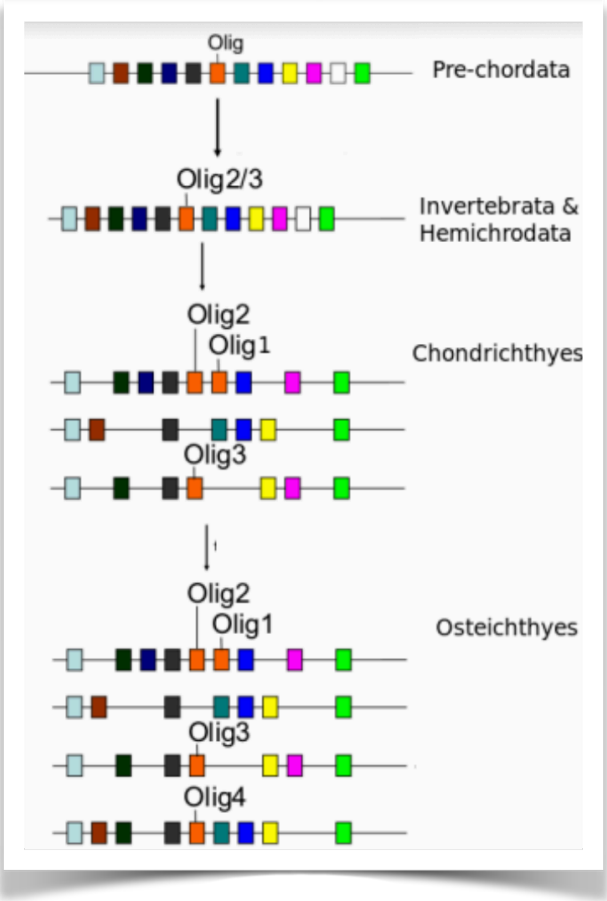


Figure 28. *Hypothetical evolution of the OLIG genes*

Finally, the hypothesis for the OLIG family of genes can be made. The hypothesis is that one proto-olig gene mutated and gained the function and with that the first OLIG2 homologue evolved. With a chromosomal duplication the homologoue of OLIG3 was created. Furthermore, with local duplication of OLIG2 the homologue of OLIG1 was created. This is appearent in Chondrichthyes which have OLIG1, 2 and 3. Afterwards, OLIG4 evolved with chromosomal duplication of OLIG3 - present in Osteichthyes and Amphibia. After this duplication, in Reptilia and Mammalia OLIG4 has been lost and in Aves both OLIG1 and 4 are not present. This absence of OLIG1 in Aves can be either a real reflection of the genome or just the result of not fully sequences genome database.

5. RNA-SEQ ANALYSIS

To analyse the bulk expression of OLIG2, the Genotype-Tissue Expression (GTEx) project database (https://www.gtexportal.org) was searched with OLIG2 gene as a query and a major expression in the brain was obtained, represented in transcripts per million (TPM). In comparison to OLIG1, it is noticable that expressions are indeed similar (*Figure 25*).

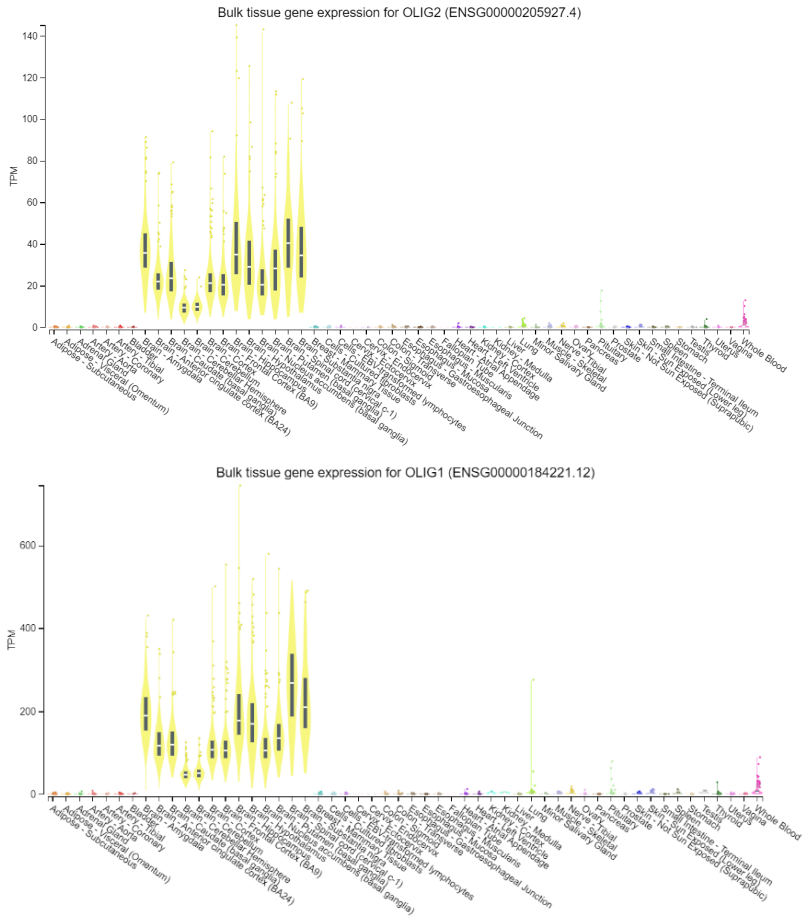


Figure 29. *The expression graphs showing the similarity in the expression sites of OLIG1 and OLIG2*

SINGLE-CELL ANALYSIS

After figuring out that OLIG2 is expressed mostly in the brain, the center of a further analysis was to find out which cell types represent the given expression within the brain tissue. A data from the scientific paper whose focused was on autism (Velmeshev et al. 2019) was used. The reason is that OLIG2 overexpression has been linked to autism disorder (Szu et al. 2021), and the experiment of Velmeshev team provided an expression matrix and metadata from cerebral cortex samples in autism patients. We supposed that having this overexpression data, the results would be more representative. After downloading the expression matrix and metadata, a Seurat object was created in R, normalized and the data scaled, and then PCA was run.

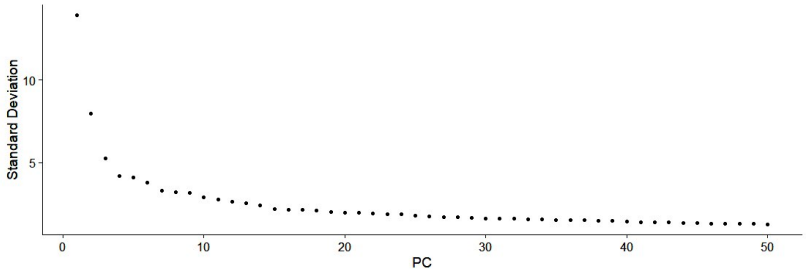
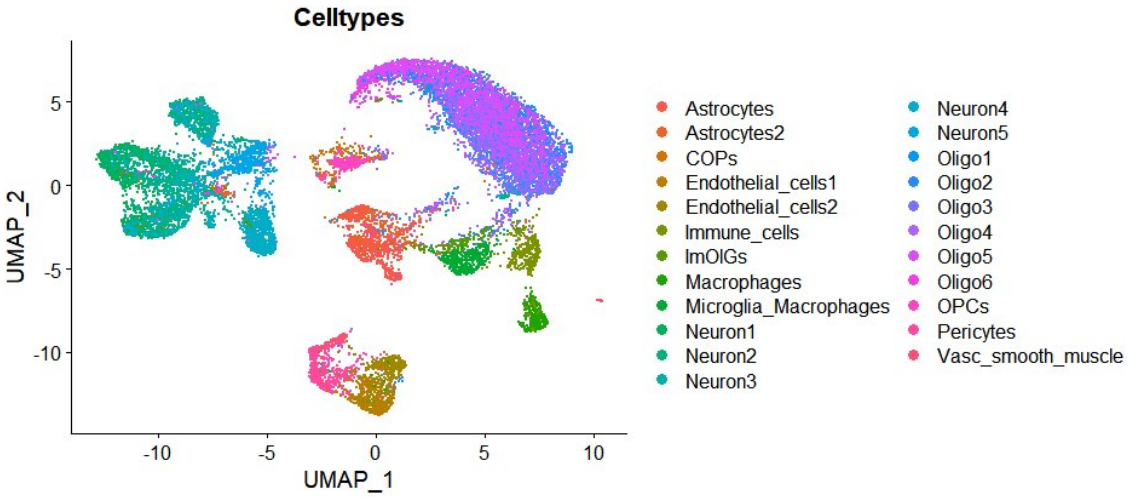


Figure 30. *PC*

Based on the graph, 14 dimensions were selected as representative from PCA. Furthuremore, neighbours were found and grouped in clusters with the Louvain algorithm. To annotate the information, we linked the information on cell types on the metadata to the Seurat object. The UMAP representation of the clusters can be seen in *Figure 27*, so we can now identify the cell types from the clusters.



Curiously, several cell types representing the same cell type were dispersed in their cluster. For instance, Oligo1-Oligo6 represent oligodendrocytes, and in the case of the oligodendrocyte precursor cells (OPCs), the name COPs represent the same. We attribute this situation to the fact that in the study we collected the data from, t-SNE dimensionality reduction was used instead of the UMAP. To have a graphical idea of this concept the *Figure 28* can be checked and then, it can be seen how these synonymous terms overlap.

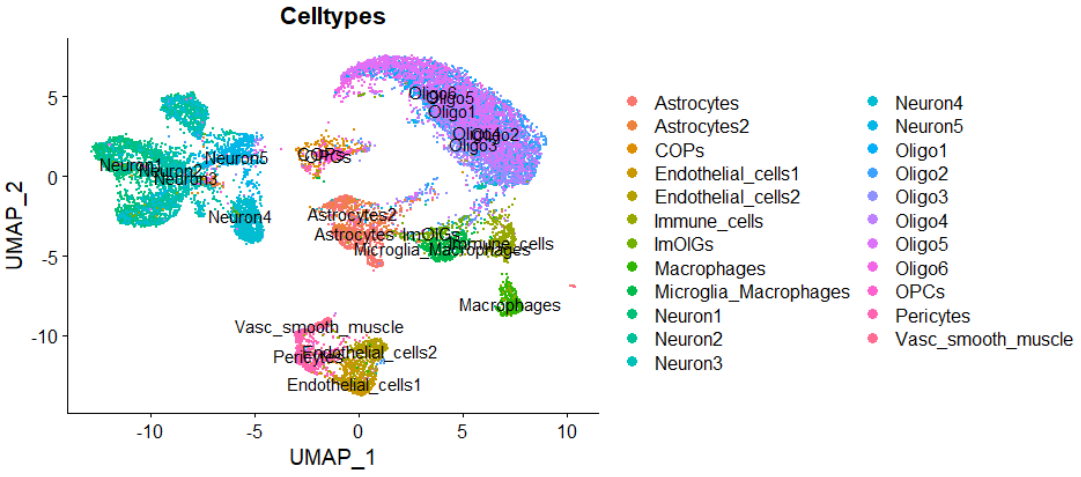


Figure 31. *UMAP cluster representation with assigned celltypes*

Finally, by focusing only on OLIG2, we see where it is expressed in the samples of the study to extrapolate it in the brain tissue. By comparing both graphs, we could conclude that the gene is expressed mainly in OPCs and OLs, but also in neurons and astrocytes.

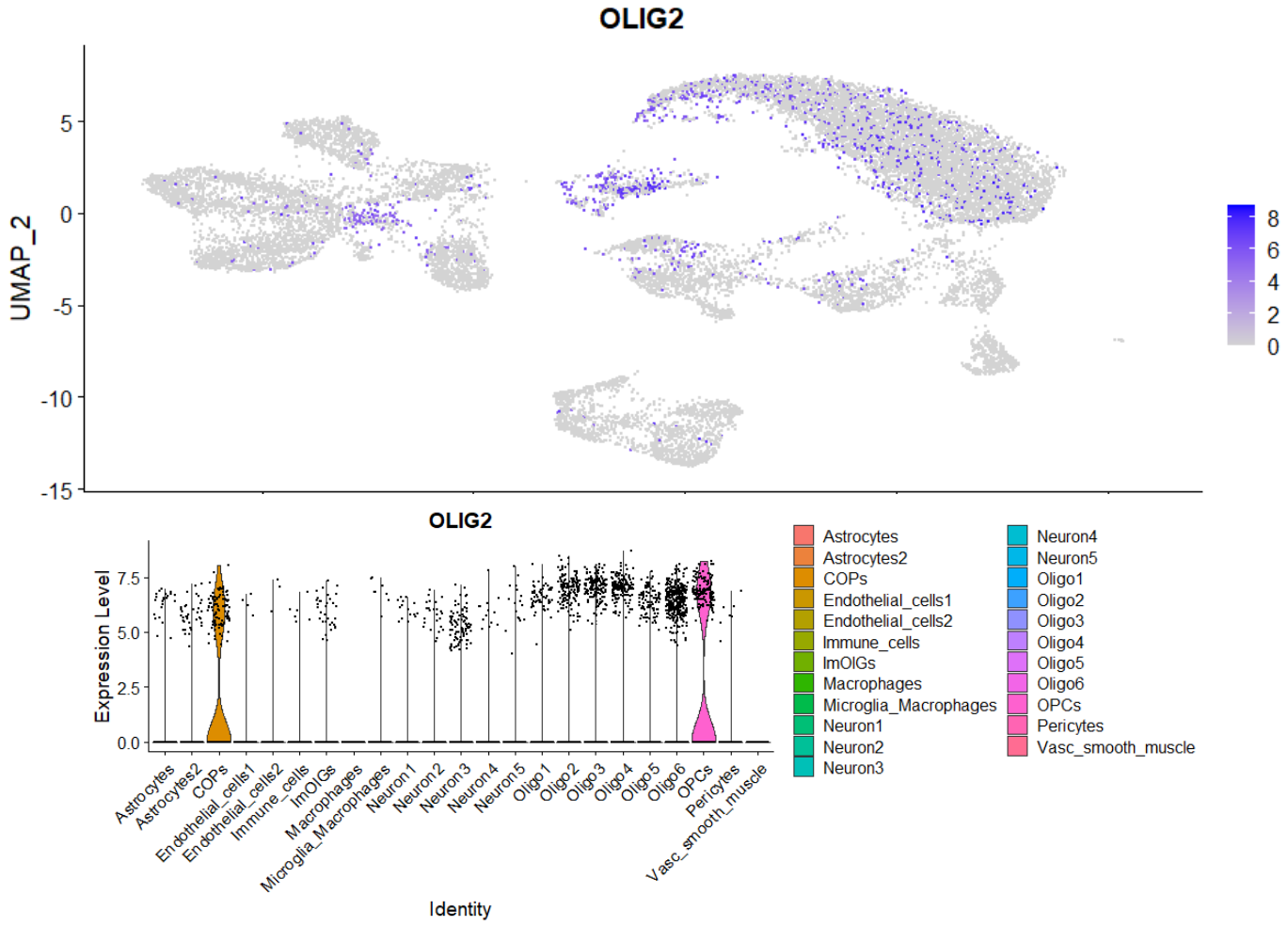


Figure 32. *Olig groups can be merged and COPs matches with OPCs, so it has been represented differently) Results obtained can be compared to the paper and in terms of expression we see that they are quite similar, the only difference is that in their research, t-SNE was used*

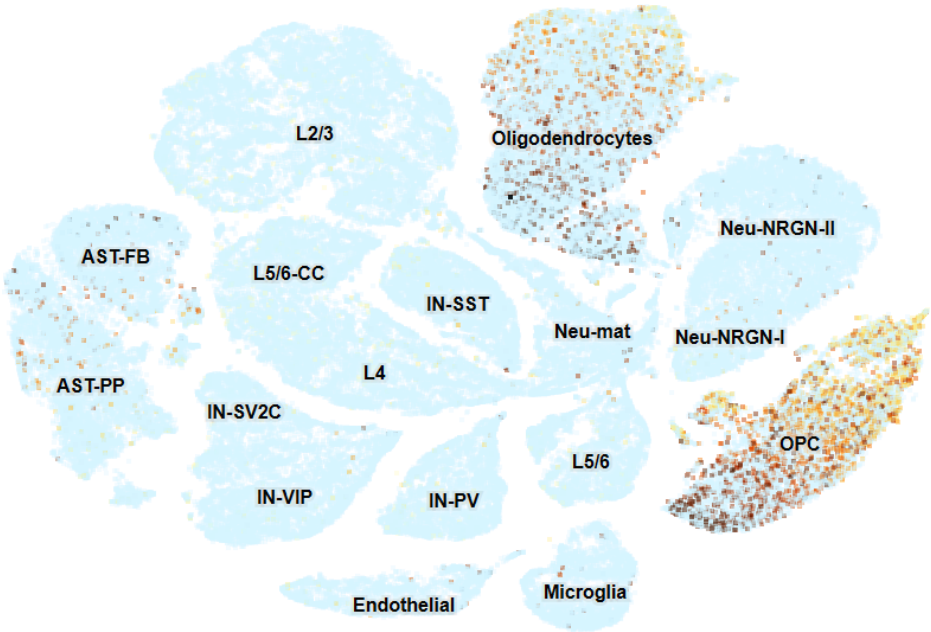


Figure 33. *Results obtained can be compared to the paper and in terms of expression we see that are quite similar [1]*

R SCRIPT USED FOR THE ANALYSIS

```
require(Seurat)
require(data.table)
library(dplyr)
library(Matrix)
library(stringr)
library(readr)
setwd("C:/Users/PC/OneDrive/Estudis/Master_BIOINFO/PGB/bHLH_project/OLIG2")

# Create SeuratObject
mat <- fread("exprMatrix.tsv.gz")
meta <- read.table("meta.tsv", header=T, sep="\t", as.is=T, row.names=1)
genes = mat[,1][[1]]
genes = gsub(".+[|]", "", genes)
mat = data.frame(mat[,-1], row.names=genes)
tiss <- CreateSeuratObject(counts = mat, project = "OLIG2", meta.data=meta)

tiss <- AddMetaData(object = tiss, meta) # Add meta data

tiss <- NormalizeData(object = tiss, scale.factor = 1e6) # Normalize data
# Computationally powerful
tiss <- ScaleData(object = tiss) #Scale data

###### We centered and scaled the data #####

# Find highly variable genes
tiss <- FindVariableFeatures(object = tiss)
tiss@assays$RNA@var.features # Now we see top variable genes
top10 <- head(VariableFeatures(tiss), 10)
plot1 <- VariableFeaturePlot(tiss)
plot2 <- LabelPoints(plot = plot1, points = top10, repel = TRUE)
plot1 + plot2 # We have a model

# Now we run the PCA analysis
#Choose a number of PCs based on the elbow plot.

tiss <- RunPCA(tiss, features = VariableFeatures(object = tiss))
ElbowPlot(tiss, ndims = 50) # We select 14 Pcs as representative

# Let's find clusters.
tiss <- FindNeighbors(tiss, dims = 1:14)
# Louvain algorithm
tiss <- FindClusters(tiss, resolution = 0.5)

## Run UMAP to visualize them:

tiss <- RunUMAP(tiss, dims = 1:14)
DimPlot(tiss, reduction = "umap", label = TRUE)

#Annotation

anno <- read.csv("meta.tsv", sep = "\t", header=T, as.is=T, row.names = 1)
tiss@meta.data$cell <- rownames(tiss@meta.data)
SetIdent(tiss, value= anno$Celltypes)
Idents(tiss)

tiss <- AddMetaData(object = tiss, anno$Celltypes, col.name = "Celltypes")
tiss@meta.data$Celltypes[is.na(tiss@meta.data$cell)] <- "unknown"
tiss$Celltypes <-as.factor(tiss$Celltypes)
```

```
# Cluster annotation (celltypes)
DimPlot(tiss, reduction = "umap", group.by = 'Celltypes', label = TRUE)

FeaturePlot (tiss, features = "OLIG2") # OLIG2 expression
VlnPlot(tiss, group.by = "Celltypes", features = c("OLIG2"))
```

6. CHIP-seq ANALYSIS

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is one of the most widely used methods which aims to determine the genomic sites that interact with a protein of interest. In this study, ChIP-seq was used to identify transcription factor OLIG2 binding sites.

High-throughput sequencing data for OLIG2 has been retrieved through international data repository NCBI Sequence Read Archive (SRA). From the repository two *.fastq* files were downloaded, one for the input and one corresponding to the peaks.

fastq-dump --split-files --verbose --gzip --outdir $PWD/01_rawData SRR1583894 (INPUT)
fastq-dump --split-files --verbose --gzip --outdir $PWD/01_rawData SRR1583890 (PEAKS)

General quality control (QC) is the first step of the ChIP-seq data analysis being performed after receiving the raw sequence files with the aim to ensure that the sequencing went well, that there was no contamination and that the library was complex enough. During many sequencing runs, the base calling accuracy declines with increasing cycles due to progressive degradation of the sequencing chemistry. This systematic quality changes is seen in the distribution of Q scores for each sequencing cycle. From the obtained results of the input file it is visible that the quality is good with a slight drop at the end. Quality control for the peaks showed more distrurbance at the end of the graph which is why the correction process was performed. Below are the graphical and tabelar representations of the quality control results without and with correction, respectfully.

SRR1583894 (INPUT RESULTS)

fastqc

w/ correction:

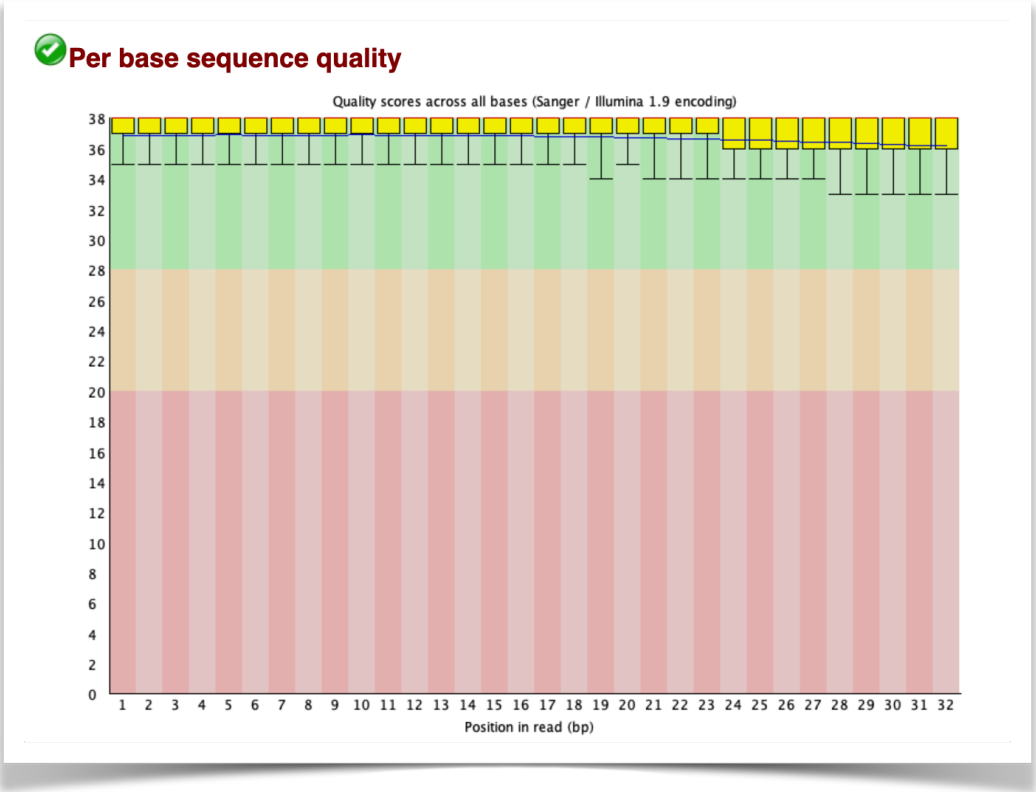


Figure 34. *FASTQC html results for the input sequences*

SRR1583890 (PEAKS' RESULTS)

fastqc

w/ correction:

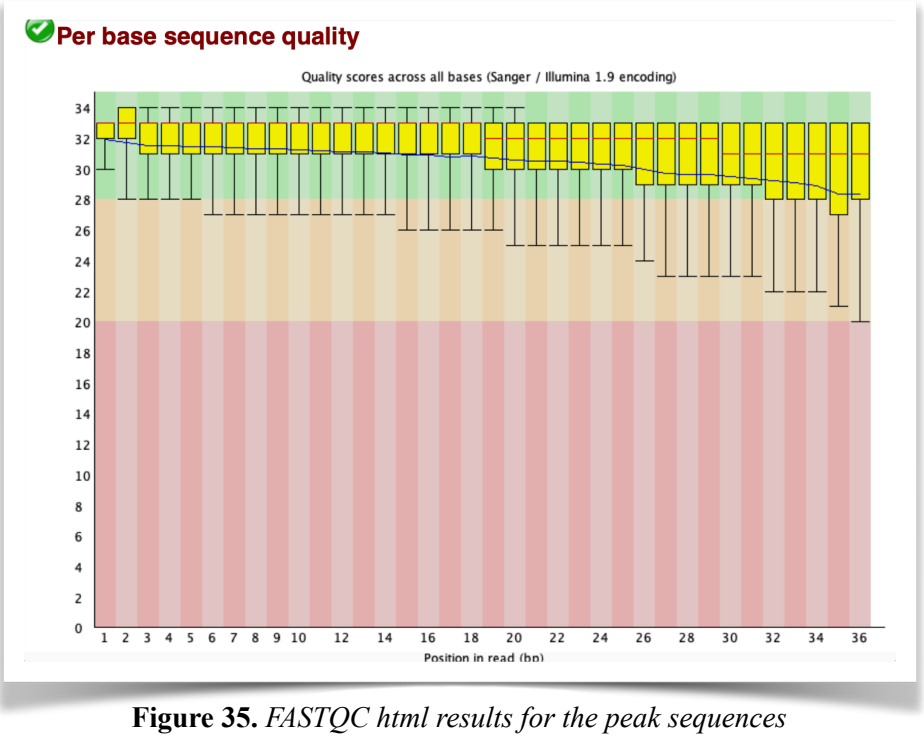


Figure 35. *FASTQC html results for the peak sequences*

MULTIQC REPORT

w/ correction:

Sample Name	% Dups	% GC	M Seqs
SRR1583890_1	52.8%	44%	29.7
SRR1583894_1	3.3%	45%	21.6

Table 2. *Multiqc report before trimming*

SRR1583890 (PEAKS' RESULTS)

fastqc

with correction:

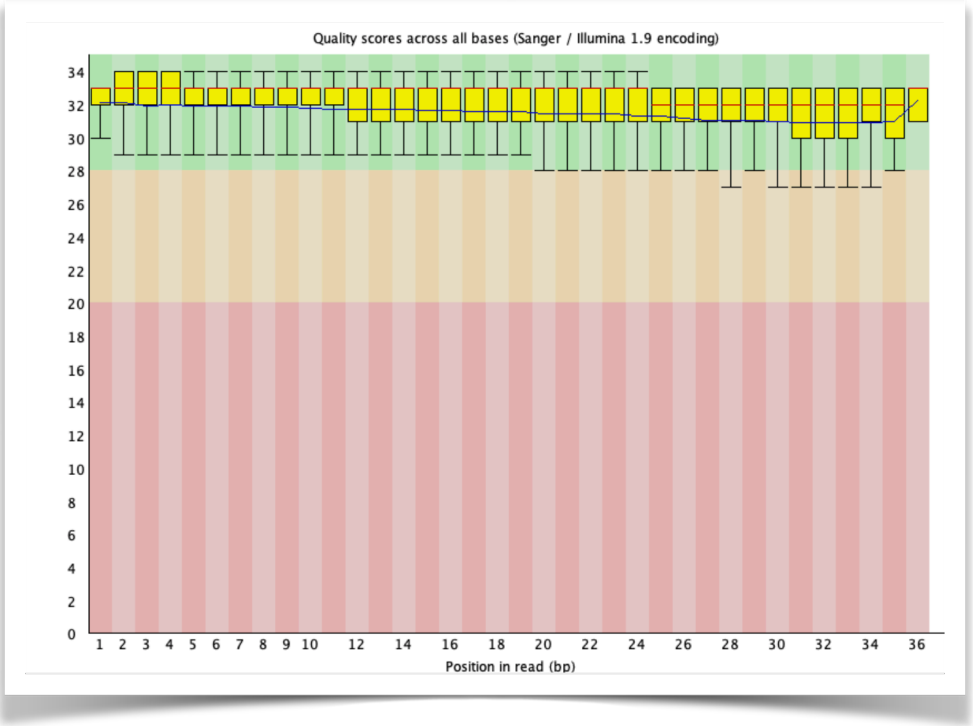


Figure 36. *FASTQC html results for the peak sequences after trimming*

MULTIQC REPORT

with correction:

Sample Name	% Dups	% GC	M Seqs
SRR1583894_trim	3.1%	45%	21.3
SRRXXX1583890_trim	38.3%	42%	24.9

Table 3. *Multiqc report after trimming*

The majority of reads originate from sequences which only occur once within the library, therefore indicating a low duplication rate or high library complexity.

Genomic alignment of the high-throughput sequencing reads to a reference genome indentifies the origin of the co-purified DNA fragments. The alignment was done using BWA algorithm after indexing the mouse genome.

```
bwa index mm10.fa
bwa mem mm10-001.fa ../../seqences_original/input_unziped/input.fastq > input.sam
bwa mem mm10-001.fa ../../02_trimmed_cutadapt/fastq_not_zipped/trim_cutadept.fastq > peaks.sam
```

BAM files have been sorted and indexed.

```
samtools sort -O bam input.sam > input.bam
samtools index input.bam
samtools sort -O bam peaks.sam > peaks.bam
samtools index peaks.bam
```

The fraction of aligned reads should be higher than 70% and ChIP-seq samples commonly show around 80% unique read start position. This is usually higher for TFs which reflects their more selective binding at distinct sites.

Peak calling represents the process of identifying the genomic regions occuped by the transcription factor of interest. The enriched DNA fragments are centered around the motif leading to sharp regions of enrichment. The process works with sliding a window along the genome, calculating an enrichment of reads in ChIP versus input samples and defining a significance score corrected for multiple testing. The sliding window usually corresponds to twice the estimated fragment size. MACS3 was used to identify OLIG2 binding sites.

```
macs3 callpeak -g mm -f BAM -t peaks.bam -c input.bam --bw 200 --outdir . -n macs_output
intersectBed -a peaks.bam -b input.bam > RobustPeaks.bed
```

Ater running MACS3 in a command line a couple of output files were obtained. Further, the obtained files were observed in IGV (*Integrative Genomics Viewer*).

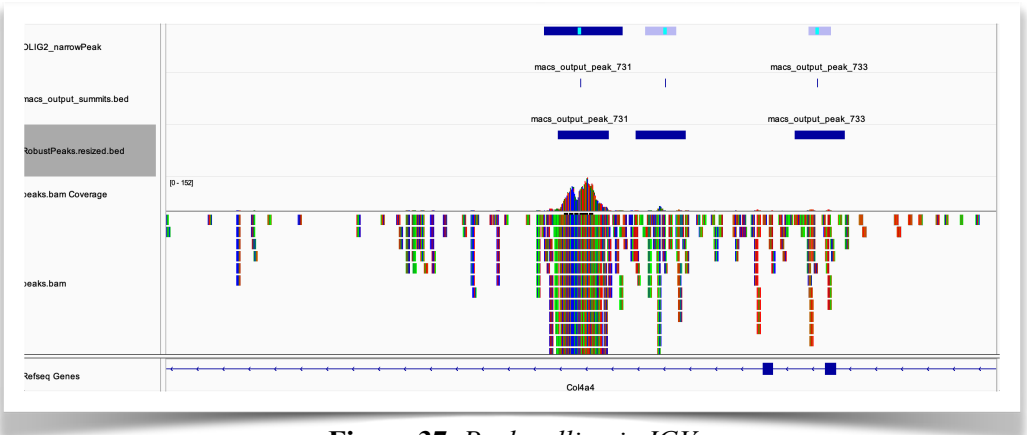


Figure 37. *Peak calling in IGV*

The number of peaks was calculated as a number of lines from the output file *macs_output_peaks.narrowPeak* and it showed that there are 22511 peaks in total. From the *.summits.bed* output file we could observe the highest point location of every peak. This file is used to find the motifs at the binding sites.

```
awk '{print $1,$2+int(($3-$2)/2)-250,$2+int(($3-$2)/2)+250}' RobustPeaks.bed | tr " " "\t" | sort -k1,1 -k2,2n > RobustPeaks.resized.bed

fastaFromBed -fi the_used_genome/mm10-001.fa -bed RobustPeaks.resized.bed > RobustPeaks.resized.bed.fa

meme-chip -oc IP_meme --db jaspar.meme RobustPeaks.resized.bed.fa
```

To find which motifs are centrally enriched in the peaks that were found, the nucleotide sequence from each peak from the genome was extracted using *fastafromBed* command and mouse genome and *Robust.Peaks.resized.bed* as the input. Afterwards, **meme-chip** was initialized and the results were observed in the browser.

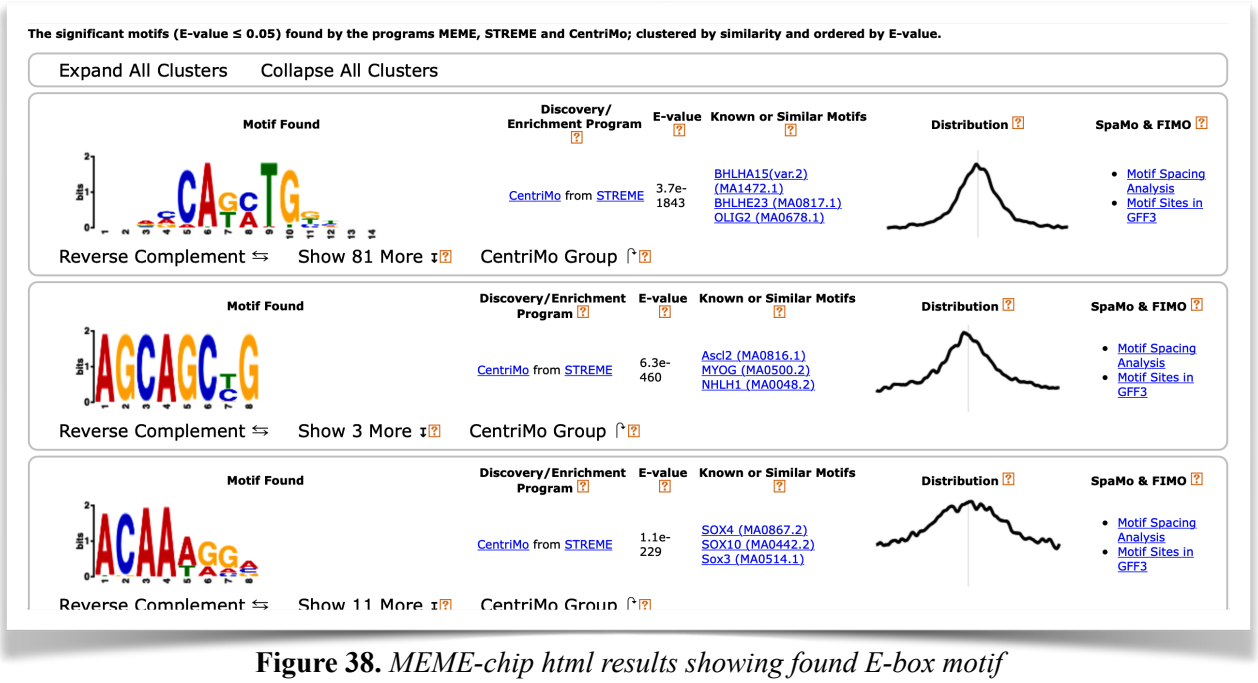


Figure 38. *MEME-chip html results showing found E-box motif*

The letter frequencies in the column are sorted in decreasing order and letters with frequency less 50% of the maximum are discarded. The results showed there is a significant motif found with an E-vaue of 3.7e-1843 with approximately normal distribution. The section next to E-value shows known or similar motifs by comparing motifs reported by a motif discovery program with known motifs in a database. It lists up to three most similar motifs and as it is shown, it recognized some of the bHLH family members, as well as OLIG2, which was our transcription factor of interest.

The graph shows the distribution of the best matches to the motif in the sequences. The sequence that is shown by CentriMo analysis is **CANNTG** which is described as an **E-box**. E-box is the palindromic sequence motif considered as the binding site for the basic helix-loop-helix class of DNA-binding proteins. Their binding specificity depends on both the nature of the "NN" nucleotides and sequences in the vicinity of the E-box. The motif is present in the regulatory elements of many developmentally controlled genes.

Different research articles point out that it is still unclear what role OLIG2 transcriptional regulation plays in neuroglia, particularly OPCs (oligodendrocyte precursor cells) and in mature OLs (oligodendrocytes) within the adult spinal cord. [1] Some research found out that OLIG2 occupies a region upstream of Mir219a-2 locus and might therefore regulate Mir219a-2 in OLs in the adult spinal cord which reveals a potentially important mechanism whereby OLIG2 regulates OL differentiation and myelination in the adult *in vivo*. As targets of OLIG2, Cdk4, Mcm4 and Mcm5 were also identified. As those enzymes are regulator of cell cycle, this suggests that in the adult spinal cord, OLIG2 continues to influence OPC proliferation via the cell cycle. [2]

7. REFERENCES

Section 1

1. Masahira N, Takebayashi H, Ono K, Watanabe K, Ding L, Furusho M, Ogawa Y, Nabeshima Y, Alvarez-Buylla A, Shimizu K, Ikenaka K. OLIG2-positive progenitors in the embryonic spinal cord give rise not only to motoneurons and oligodendrocytes, but also to a subset of astrocytes and ependymal cells

Section 2

1. Conserved Domains Database (CDD) and Resources [Internet]. [cited 2022 Nov 26]. Available from: https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

2. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 2012;8(7):e1002764
3. OLIG2 - Oligodendrocyte transcription factor 2 - Homo sapiens (Human) | UniProtKB | UniProt [Internet]. [cited 2022 Nov 26]. Available from: https://www.uniprot.org/uniprotkb/Q13516/entry

4. SLAC · Issue #1425 · veg/hyphy [Internet]. GitHub. [cited 2022 Nov 26]. Available from: https://github.com/veg/hyphy/issues/1425

5. The Population Genetics of dN/dS - PMC [Internet]. [cited 2022 Nov 26]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596312/

6. KVEG. HyPhy - Hypothesis Testing using Phylogenies [Internet]. [cited 2022 Nov 26]. Available from: http://veg.github.io/hyphy-site/

Section 3 and 4

1. Atchley WR, Fitch WM. A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci U S A. 1997 May 13;94(10):5172-6. doi: 10.1073/pnas.94.10.5172. 9144210; PMCID: PMC24651.

2. Darr AJ, Danzi MC, Brady L, Emig-Agius D, Hackett A, Golshani R, et al. (2017) Identification of genome-wide targets of OLIG2 in the adult mouse spinal cord using ChIP-Seq. PLoS ONE 12(10): e0186091

3. De Martin X, Sodaei R, Santpere G. Mechanisms of Binding Specificity among bHLH Transcription Factors. Int J Mol Sci. 2021 Aug 24;22(17):9150. doi: 10.3390/ijms22179150. PMID: 34502060; PMCID: PMC8431614.

4. Dev Biol. 2006 May 15;293(2):358-69. doi: 10.1016/j.ydbio.2006.02.029. Epub 2006 Apr 3. PMID:16581057.Li H, Richardson WD. Evolution of the CNS myelin gene regulatory program. Brain Res. 2016 Jun 15;1641(Pt A):111-121. doi: 10.1016/j.brainres.2015.10.013. Epub 2015 Oct 22. PMID: 26474911; PMCID: PMC6326354.

5. Ledent V, Vervoort M. The basic helix-loop-helix protein family: comparative genomics and

phylogenetic analysis. Genome Res. 2001 May;11(5):754-70. doi: 10.1101/gr.177001. PMID: 11337472;PMCID: PMC31104

6. Li H, Richardson WD. The evolution of Olig genes and their roles in myelination. Neuron Glia Biol.2008 May;4(2):129-35. doi: 10.1017/S1740925X09990251. PMID: 19737433; PMCID: PMC6326352.

8. Meijer DH, Kane MF, Mehta S, Liu H, Harrington E, Taylor CM, Stiles CD, Rowitch DH. Separated at birth? The functional and molecular divergence of OLIG1 and OLIG2. Nat Rev Neurosci. 2012Dec;13(12):819-31. doi: 10.1038/nrn3386. PMID: 23165259; PMCID: PMC3733228.

Section 5

1. Szu, J., Wojcinski, A., Jiang, P., & Kesari, S. (2021). Impact of the Olig Family on Neurodevelopmental Disorders. In Frontiers in Neuroscience (Vol. 15). https://doi.org/10.3389/fnins.2021.659601

2. Velmeshev, D., Schirmer, L., Jung, D., Haeussler, M., Perez, Y., Mayer, S., Bhaduri, A., Goyal, N., Rowitch, D. H., & Kriegstein, A. R. (2019). Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, *364*(6441). https://doi.org/10.1126/science.aav8130

Section 6

1. Terme, JM., Calvignac, S., Duc Dodon, M. *et al.* E box motifs as mediators of proviral latency of human retroviruses. *Retrovirology* 6, 81 (2009)

2. Yutzey KE, Konieczny SF. Different E-box regulatory sequences are functionally distinct when placed within the context of the troponin I enhancer. Nucleic Acids Res. 1992 Oct 11;20(19):5105-13. doi: 10.1093/nar/20.19.5105. PMID: 1329039; PMCID: PMC334291